Bias Analysis for Logistic Regression with a Misclassified Multi-categorical Exposure

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Collaborative Graduate Program in Biostatistics
University of Saskatchewan
Saskatoon

By
Yaqing Liu

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Executive Director, School of Public Health,
107 Wiggins Road,
University of Saskatchewan,
Saskatoon, Saskatchewan
Canada
S7N 5E6

Abstract

In epidemiological studies, it is one common issue that the collected data may not be perfect due to technical and/or financial difficulties in reality. It is well known that ignoring such imperfections may lead to misleading inference results (e.g., fail to detect the actual association between two variables). Davidov et al.(2003) have studied asymptotic biases caused by misclassification in a binary exposure in a logistic regression context. The aim of this thesis is to extend the work of Davidov et al. to a multi-categorical scenario. I examine asymptotic biases on regression coefficients of a logistic regression model when the multi-categorical exposure is subject to misclassification. The asymptotic results may provide insight guide for large scale studies when considering whether bias corrections would be necessary. To better understand the asymptotic results, I also conduct some numerical examples and simulation studies.

Acknowledgements

In the first place I would like to show my gratitude to my supervisor, Juxin Liu, Ph.D., whose encouragement, guidance and advice from the very early stage of this thesis to the final level enabled me to develop an understanding of the subject.

I am also indebted to Professor Mik Bickis, Ph.D., Professor Lisa Lix, Ph.D., Professor Longhai Li, Ph.D., and Professor Bonnie Janzen, Ph.D., who directed me in biostatistics. Their kind support and guidance of my education have been of great value in this thesis.

I am grateful many of my colleagues, my friends and family for their constant belief in me, and major support during my life.

Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of this thesis.

Contents

List of Tables

Chapter 1

Introduction

In epidemiological studies, one common issue is that errors may contaminate the assessment of the exposure, where an exposure refers to the variable of interest that alters individual's risk of developing a certain disease[1]. The scenario of an exposure mismeasured is referred to as involving *mismeasurement*[2]. More specifically, the term *measurement error* often refers to mismeasurement in a continuous variable, and the term *misclassification* often refers to mismeasurement in a categorical or discrete variable[3, 4, 5, 6, 7]. Researchers have shown that naively ignoring mismeasurement may induce biased estimates with falsely small standard errors[7]. Under the regression context, the predictivity of the covariate(s) on the response variable may also be reduced by mismeasurement in the covariate(s)[8]. Drews et al.[9] further state misclassification in exposure is a more serious issue than measurement error because misclassification could result in a high number of biases when even a small number of errors occurs.

The aim of this thesis is to extend the work of Davidov et al.[10] from a binary case to a multi-categorical (i.e. more than two categories) scenario. Asymptotic bias formulas will be derived to assess biases caused by misclassification in a multi-categorical exposure in a logistic regression context. The asymptotic results of this study provide a way to quantify the biases for studies with large sample sizes. Based on these quantities, researchers may decide whether correction methods should be applied. In other words, researchers can use the asymptotic results to obtain some idea whether misclassification errors are harmful enough and thus need to be taken care of.

The outline of this thesis is as follows. Chapter 1 describes some possible sources of mis-

measurement and introduces the mismeasurement issue along with associated terminology. A literature review about misclassification is given in Chapter 2. In Chapter 3, I derive the asymptotic bias formulas when only one misclassified multi-categorical exposure is in the model and when one error-free binary covariate is also included in the model. Chapter 4 illustrates the proposed methodology through numerical instances. Moreover, simulation studies for finite sample cases are also conducted to compare with the results from numerical examples. All results calculated by R-software are shown in Appendix A and Appendix B. Finally, Chapter 5 gives a summary, points out the significance of my work, and states opportunities for future researches.

1.1 Sources of Mismeasurement

Various errors could occur when we collect the information either from people or about people. Generally speaking, errors can be separated into two categories: *random error* and *systematic error*[4, 11, 12].

1.1.1 Random Error

A random error stands for the error that occurs simply by chance. If we denote E as the true value of the exposure and X as the observed value of E (i.e. surrogate), then random errors result in $\mathrm{E}(X) = \mathrm{E}(E)$ and $\mathrm{V}(X) > \mathrm{V}(E)$; that is, X has the same expectation as E but has a larger variance. Random errors in a continuous variable may not cause a pressing concern to investigators due to the null expected value it has[11, 13]. However, it does matter when random errors take place in a categorical (i.e. nominal or ordinal) exposure since there is no "buffer zone" in a sense that the information is distributed in either a correct category or an absolutely wrong category[11]. On the other hand, if researchers decide to convert a continuous variable into a discrete one, then random errors may become influential as a result of the measurement error problem probably turning into a misclassification scenario[9, 11]. For instance, assume a group of individuals is randomly selected based on their weight. Also assume only random errors are introduced in this sample and measurement error does not result in a biased point estimate of weight. Suppose all subjects are classified

into two categories according to their weight, then, for instance, normal-weight individuals may be wrongly labeled as overweight and vice versa (i.e. misclassification). Consequently, an inevitable bias is presumable to be present in the analysis[9, 11].

In order to reduce random errors, researchers mainly apply two methods in the design stage. The first one is to utilize the *precise* instrument that gives little variation among observations when repeating using it for the same experiment under the same circumstance[11]. In contrast, the *accurate* device refers to the one reducing systematic errors[11]. The inaccurate instrument yields the proxy value either smaller or greater than the true value (more in section 1.1.2)[11]. The second way to reduce random errors is to employ a repeated measures design[11, 14]. The precise tool and the repeated measures design will be discussed in more detail in the following two paragraphs.

The *precision* of a measurement reflects how well a device performs over time[1, 11]. Therefore, implementing a precise means is ideal to reduce the variance of the random error. Nonetheless, we have to admit that the random error always exists because there is no absolutely precise instrument. On the other hand, although we are aware it is crucial to use the precise tool, technical and financial difficulties sometimes force researchers to carry out the indirect measure, which gives erroneous data with a large variance, to gather the information. For instance, the Radiation Effects Research Foundation (RERF) in Hiroshima conducted a study to assess the relationship between atomic-bomb survivors' radiation doses and radiation-related disease. The study was conducted five years after the bombs fell. The technical limitation made it unrealistic to gauge the radiation that survivors had encountered using the direct instrument, so the indirect factors such as location and shielding were employed to estimate radiation doses[15]. In such a context, the observed data will differ from the actual radiation doses and have a larger variance because the imprecise indirect instrument is used. Here is another real-world instance that applies the indirect measure. A Nurses' Health Study was conducted to investigate how nutritional intakes relate to breast cancer. It was financially impractical and technically infeasible to directly measure the nutritional intakes; therefore, a self-report *food-frequency questionnaire* (FFQ) was employed to collect

the frequencies of subjects' recent specific consumed foods. Mathematical formulas were also employed to convert gathered frequencies of foods into the nutrient level acquired[16]. It turns out the transformed data are different from the actual nutritional intakes with a larger variance.

The repeated measures design is another crucial approach adopted in practice to reduce the random error and to improve the precision of measurement. The repetition can be completed through three primary ways: different diagnostics is used for every patient; the same test is used several times on each subject; and the same instrument is carried out by different conductors for the same individual[14]. The Nurses' Health Study mentioned above is an instance of repeated measurements. FFQ was used four times for every participant and each time one-week long diet intakes were recorded. The average value of the observed diet intakes were then employed in the calculation to improve the study's precision[16]. Carroll et al.[7] suggest employing a repeated measures design when the surrogate mean is believed to be closer to the true value than a single observation. Moreover, they state it is appropriate to use the classical error model (see section 1.2.2) when repeated measurements are employed. Walter and Irwig[14] claim the estimation of *odds ratio*(see section 1.3.1) in a case-control study will be substantially improved even if repeated measurements are applied to only a portion of subjects who are subject to misclassification. A case-control study refers to a study in which the disease status of each participant is recorded at the baseline of the study. So that the exposure value is recalled to escape the follow-up period after individual is exposed to the variable of interest[11, 17].

1.1.2 Systematic Error

The error that develops in a systematic way is named as systematic error or systematic bias[11, 12]. It is a more serious problem compared to random error in mismeasurement studies, since the poor *accuracy* induced by systematic biases results in $\mathrm{E}(X) \neq \mathrm{E}(E)$, where the accuracy describes the closeness of the measurement to the true value[11, 4, 1]. The mathematical equation $\mathrm{E}(X) \neq \mathrm{E}(E)$ indicates the expectation of the surrogate variable does not

equal the expectation of the true variable. Systematic errors can be reduced by using the ideal instrument to increase the accuracy, yet it is possible no accurate instrument is available. Limited resources (eg. funding and time) could also make it unlikely to implement the accurate measure on every individual in the sample. The following paragraph shows a way of how investigators coping with such a situation in practice by employing a real world example.

A study in Saarland, Germany investigated the relationship between stomach cancer and *Helicobacter pylori* (*H.pylori*) infection. *H.pylori* is one kind of bacterium that can inhabit the stomach[18]. A case-control study was conducted in order to avoid the possible high expenses resulting from the long waiting of having stomach cancer after the infection. Besides, in spite of sacrificing the accuracy, an easiest and cheapest way (i.e. a laboratory test) was employed to check the presence of *H.pylori* infection. On the other hand, investigators applied the accurate measurement (i.e. gold standard, see section 1.3.3) to a selected subgroup of sample (i.e. an internal validation group, see section 1.3.3). In such a way, researchers have not only proxy data but also true data for the validation group, which can be used to correct misclassification of *H.pylori* infection in the remaining sample where researchers have only the contaminated data (i.e. main sample, see section 1.3.3)[18].

Systematic errors can be introduced by various ways. Here I describe two of them: recall bias and interviewer bias.

Recall Bias

Recall bias is a type of bias that takes place in a retrospective process. Retrospective studies are those looking backwards in time; that is, the outcome is observed before the covariate[19]. Case-control study is one common type of retrospective studies, where cases represent those individuals with a particular disease and controls represent those without[12, 19]. In a case-control study, recall bias can be explained as a result of cases intentionally underestimating or overestimating their exposure while controls not[12, 19]. For instance, a case-control study is designed to examine the relationship between certain respiratory disease and passive smok-

ing. A recall bias may arise during the recall procedure as cases intentionally exaggerate the smoking amount they have been exposed if they think the amount of passive smoking is the reason for them to have the disease. Controls do not intentionally do so. As a result, the exposure risk will be overestimated for cases[11].

One issue worth pointing out is the differences between recall biases and recall errors. Recall errors are those unescapable errors for both cases and controls due to the memory failure of individuals[11]. All case-control studies (e.g., nutrition intake and *H.pylori* instances described above) experience some degree of recall errors and the number of errors is expected to be same for cases and controls. Recall errors are very likely related to the non-differential mismeasurement (refer to section 1.2.3)[11]. In contrast, recall biases do not generally present in all case-control studies. Furthermore, recall biases tend to result in differential mismeasurement (see section 1.2.3) as a consequence of unequal number of errors in cases and controls[11].

Interviewer Bias

Interviewer bias is one type of biases resulting in systematic error[11]. For instance, in a Massachusetts Women's Health Study, 2,569 women were selected to conduct a 5-year long cohort study. A prospective cohort study is a study following the natural way so that the exposure is measured before the response variable. Six interviewers were assigned to phone each participant to finish an approximately 30-minute long questionnaire and each subject received six calls total through the five-year period. Johannes et al.[20] found there was interviewer variation for questions in need of further probing or questions relating to subjective or personal topics. The authors further claimed, despite the interviewers receiving the same training, the descriptions the interviewers entered were different due to different recording ways of them even though they heard the same answer from the same interviewee. The variation of different interviewers finally results in an interviewer bias in recording the information.

1.2 Measurement Error and Misclassification Models

This section describes terminology for modeling mismeasurement. For a more systematic discussion of measurement error and misclassification, please refer to Fuller[21] and Carroll et al.[7].

1.2.1 Functional Models versus Structural Models

The property of the unobserved true exposure E can be treated as a standard to broadly classify mismeasurement models[2, 7, 3, 22]. That is, if E is fixed, we say the mismeasurement model is *functional* and apply a functional modeling approach to account for possible mismeasurement of E[7, 22]. Carroll et al.[7] have suggested a more fruitful definition for the functional model. That is, one model is also functional if E is a random variable but minimal assumptions have made about it. In other words, this definition includes the nonparametric situation. By contrast, if E is random and parametric assumptions are made on the distribution of E, then the model is named as *structural* model[7, 4, 3, 22].

So far, it is unclear in the literature whether functional or structural modeling is more preferable in general[7]. Carroll et al.[7] have suggested using the functional model when the estimator obtained by using the functional modeling is distributional-robust (i.e. the estimator is consistent). The functional modeling also takes advantage of making less assumption on the unobserved E compared to the structural modeling. On the other hand, structural methods are more attractive because they are usually more efficient compared to functional methods, where the efficiency is gained through making parametric assumptions on the distribution of E[7].

1.2.2 Models for Measurement Error and Misclassification

We can also separate mismeasurement models into *classical error models* and *Berkson error models* based on the relationship between the true but unobserved E and its surrogate X [7, 3, 23]. Classical error models represent those making assumptions on the distribution of X

given E and the error-free covariate(s) Z. While Berkson error models are models concerning the distribution of E conditional on (X,Z)[7, 3, 23]. In the following, I describe classical error models and Berkson error models in a mathematical way.

Classical Error Models

The classical error model can be written as follows:

$$X = E + U,$$

where U is independent of E and has mean zero. The model indicates X is unbiased for E, i.e. $E(X|E, Z) = E$ or $E(U|E, Z) = 0$. X and U are two random variables with either fixed or random variances[7, 3].

In reality, we have to admit the proxy X could be biased from the true E when the systematic bias occurs. Therefore, we need to calibrate the biased measurement resulting in an unbiased measurement, which gives the error model as below:

$$X = \alpha_0 + \alpha_1 E + \alpha_2 Z + U,$$

where U is independent of E and Z and E($U|E, Z$)=0. This model says $(\frac{X-\alpha_0-\alpha_2 Z}{\alpha_1})$ is an unbiased measure of E. Again variances of random variables U and X can be either fixed or random[7, 3].

Berkson Error Models

Different from a classical error model, in a Berkson error model, the true E is influenced by X and Z. For instance, a herbicide study conducted by Rudemo et al.[24] employed the Berkson error model. The exposure of interest is the actual amount of herbicide absorbed by the plant, which is unmeasurable. The amount of herbicide required to spray to a plant is fixed and measurable, but not equivalent to the absorbed concentration. In such a context, the true value of the herbicide concentration depends on the surrogate amount. Therefore, a Berkson relationship is appropriate. The Berkson error model can be written as:

$$E = \rho_0 + \rho_1 X + \rho_2 Z + U^*,$$

where U^* is independent of X and Z and E($U^*|X, Z$)=0. The model indicates ($\frac{E-\rho_0-\rho_2 Z}{\rho_1}$) is an unbiased estimator of X. The variance of U^* can be either fixed or random.

1.2.3 Non-differential versus Differential Mismeasurement

Mismeasurement can be separated into two types, namely *non-differential* and *differential* mismeasurement. Non-differential mismeasurement presents if the observed exposure has no additional information about the response variable when the true value of that exposure is given[7, 2]. If the response variable is binary (e.g., disease or non-disease) and the exposure is categorical, then the non-differential circumstance simply means possible misclassification probabilities do not vary between cases and controls. Mathematically, we explain non-differential misclassification as $P(Y|E, X, Z) = P(Y|E, Z)$, where $P(A|B)$ indicates the conditional probability of A given B and Z represents all other error-free covariates[7, 4, 23]. Non-differential misclassification could be caused by random errors such as fallible memory and misunderstanding questions, or systematic errors such as test failures as long as errors are equally likely to occur in all levels of Y[11]. From the design perspective, cohort studies by nature are more likely to relate to non-differential mismeasurement than differential mismeasurement, because subjects are not aware of their future disease statuses and they are unable to alter their exposure values based on the unknown disease statuses. It is highly probable for patients to make the same number of errors in exposure[5]. One numerical instance reflects effects of ignoring non-differential misclassification shown in section 1.3.1.

In contrast, differential mismeasurement indicates the erroneous exposure has additional information about the outcome conditional on the information contained in the true exposure. Therefore, differential misclassification implies different response groups receive different number of errors. A more technical definition of differential misclassification is $P(Y|E, X, Z) \neq P(Y|E, Z)$[7, 4, 23]. Differential mismeasurement is mainly caused by systematic errors[11]. However, it could still be resulted by random errors although rare[11]. For instance, if we adopt different instruments with different precisions in cases and controls, then random error but differential mismeasurement probably arise. By contrast with cohort studies, case-control studies are prone to encounter differential mismeasurement. As I men-

tioned in the recall bias section, cases have a higher probability than controls to overestimate or to underestimate their exposure(s). Flegal et al.[5] state a non-differential case may turn into a differential scenario when continuous data that are subject to measurement error are turned into categorical data. They further point out this situation is very likely to occur even in a cohort study. One numerical instance shows influences of differential misclassification on parameters given in section 1.3.1.

1.3 Terminology Used in Epidemiological Literature

This section discusses the terminology that is often used in the epidemiological literature. They are measures of association (section 1.3.1), misclassification rates (section 1.3.2), and gold standard and validation (section 1.3.3).

1.3.1 Odds Ratio and Relative Risk

Odds Ratio

Odds ratio (OR) is one measure used in epidemiological studies to investigate the association between the exposure and the corresponding disease[19]. OR is the ratio of the odds of disease among exposed people to the odds of disease among unexposed. *Odds* represent the ratio of the probability of an event to its complement. Therefore, the odds of disease among exposed is the ratio of the probability of having disease given exposed (i.e. P(disease | exposed)) dividing its complement (i.e. 1-P(disease | exposed)=P(disease-free | exposed)). Consequently, OR is:

$$\begin{aligned} OR &= \frac{P(\text{disease} \mid \text{exposed})/P(\text{disease-free} \mid \text{exposed})}{P(\text{disease} \mid \text{unexposed})/P(\text{disease-free} \mid \text{unexposed})} \\ &= \frac{P(\text{disease} \mid \text{exposed})P(\text{disease-free} \mid \text{unexposed})}{P(\text{disease} \mid \text{unexposed})P(\text{disease-free} \mid \text{exposed})}. \end{aligned}$$

An artificial instance is used in the following to calculate OR numerically. Suppose that a case-control study is conducted to examine the relationship between smoking habit and lung cancer. Let S denote smokers, NS denote nonsmokers, D denote having lung cancer, and $\overline{D}$ denote not having lung cancer. Assume 600 cases and 550 controls are pre-determined by

Table 1.1: The frequencies of a hypothetical case-control study

	Lung Cancer	No Lung Cancer	Total
Smokers	400	300	700
Nonsmokers	200	250	450
Total	600	550	1150

researchers in advance. The frequencies of four combinations of smoking status and disease status are illustrated in Table 1.1 (a 2× 2 table).
OR can be calculated as follows:

$$\begin{aligned} OR &= \frac{P(D \bigcap S) * P(\overline{D} \bigcap NS)}{P(D \bigcap NS) * P(\overline{D} \bigcap S)} \\ &= \frac{400 * 250}{300 * 200} \\ &= 1.67. \end{aligned}$$

The numerical result of OR indicates the odds of lung cancer are 1.67 times higher among smokers compared to nonsmokers.

Note that OR can also be expressed as the ratio of the odds of exposed among cases compared to the odds of exposed among controls. Thus, OR in the above instance can be re-interpreted as the odds of smokers 1.67 times greater among those with lung cancer compared to those without. An intuitive implication of this interpretation is the equivalency of odds ratios in case-control and cohort studies. That is:

$$\begin{aligned} OR &= \frac{P(\text{disease} \mid \text{exposed})/P(\text{disease-free} \mid \text{exposed})}{P(\text{disease} \mid \text{unexposed})/P(\text{disease-free} \mid \text{unexposed})} \\ &= \frac{P(\text{disease} \bigcap \text{exposed})/P(\text{disease-free} \bigcap \text{exposed})}{P(\text{disease} \bigcap \text{unexposed})/P(\text{disease-free} \bigcap \text{unexposed})} \\ &= \frac{P(\text{exposed} \mid \text{disease})/P(\text{unexposed} \mid \text{disease})}{P(\text{exposed} \mid \text{disease-free})/P(\text{unexposed} \mid \text{disease-free})}. \end{aligned}$$

The first line of the above derivation indicates the expression of OR for a cohort study and

the third line implies the expression for a case-control study.

In the numerical instance, the exposure (i.e. smoking status) is treated as a risk factor since OR is larger than 1, which indicates a positive association between the exposure and the disease (i.e. the risk effect)[19]. If we have OR smaller than 1, then the exposure is said to have a protective effect on the disease and that exposure is treated as a protective factor[19]. Obviously, OR equivalent to 1 indicates no association. In general, the more OR departs from 1, the stronger the association between the exposure and the outcome has.

Another thing worth mentioning is the coefficients in a logistic regression can be related to odds ratios by the exponential function. More details will be discussed in Chapter 3.

Relative Risk

Besides OR, *relative risk* (RR) is another association measure based on ratios. Note that it is not appropriate to use RR for a case-control design as a result of the nature that the number of cases and controls are pre-determined in a case-control study. In other words, RR is useless if we do not have the knowledge of the total number of cases and controls in a population or proportions of cases and controls in that population[19]. But RR is still useful as an association measure for cohort studies.

RR compares the probability of disease among exposed individuals dividing the probability of disease among those unexposed. We then have:

$$RR = \frac{P(\text{disease} \mid \text{exposed})}{P(\text{disease} \mid \text{unexposed})}.$$

If we assume the data in Table 1.1 are from a cohort study, we then have RR as below:

$$\begin{aligned} RR &= \frac{P(D \mid S)}{P(D \mid NS)} \\ &= \frac{400/700}{200/450} \\ &= 1.29. \end{aligned}$$

The result indicates smokers are 1.29 times more likely to develop lung cancer than nonsmokers. In other words, the risk of developing lung cancer is 0.29 higher for a smoker than a nonsmoker.

Particularly, when the disease is rare, RR approximately equals OR as the total number of exposed is about the number of exposed but without disease (i.e. P(disease-free |exposed) $\approx$ 1) and the total number of unexposed is about those unexposed controls (i.e. P(disease-free |unexposed) $\approx$ 1)[19]. That is:

$$\begin{aligned} OR &= \frac{P(\text{disease} \mid \text{exposed})P(\text{disease-free} \mid \text{unexposed})}{P(\text{disease} \mid \text{unexposed})P(\text{disease-free} \mid \text{exposed})} \\ &\approx \frac{P(\text{disease} \mid \text{exposed}) * 1}{P(\text{disease} \mid \text{unexposed}) * 1} \\ &= RR \end{aligned}$$

1.3.2 Misclassification Rates

When there is a categorical exposure subject to misclassification, probabilities of its observed status given its true status are misclassification rates. The accuracy of a measurement tool can be reflected by misclassification rates. In other words, they are measures of the magnitude of misclassification. Misclassification rates can further help investigators correct misclassification. They can be written into a matrix form with diagonal ones representing correctly distributed probabilities (i.e. classification probabilities) and off-diagonal ones indicating misclassification probabilities. In section 3.1.1, I demonstrate the general matrix form of misclassification rates and give a detailed description.

For a simple case with only one binary exposure and one binary outcome as variables (i.e. a 2×2 case), misclassification rates include four cells, which are *sensitivity*, 1-sensitivity, *specificity*, and 1-specificity. The sensitivity of a test represents the percentage of people grouped into the exposed category given they are truly exposed. The specificity of a test indicates the proportion of those truly unexposed individuals. Sensitivity and specificity belong to classification probabilities. 1-sensitivity is the complement of sensitivity called as *false neg-*

Table 1.2: The effects of non-differential misclassification when 5% of smokers are misclassified as nonsmokers and 8% of nonsmokers are misclassified as smokers

	Lung Cancer	No Lung Cancer	Total
Smokers	400-20+16=396	300-15+20=305	701
Nonsmokers	200-16+20=204	250-20+15=245	449
Total	600	550	1150

ative rate and 1-specificity is the complement of specificity named as *false positive rate*[19]. Both of them are misclassification probabilities. For non-differential misclassification, we have one pair of sensitivity and specificity as a result of the same amount of misclassification cases and controls have. In contrast, we have two pairs of sensitivity and specificity for cases and controls separately in a differential misclassification scenario. In addition to sensitivity and specificity, researchers sometimes need probabilities to have an "inverse" form of them. In other words, the probabilities of the true status given the diagnosed one are needed[25]. Specifically, the proportion of individuals who are diagnosed as exposed truly exposed refers to *positive predictive value* (PPV) and the probability of unexposed individuals given they are identified as unexposed refers to *negative predictive value* (NPV)[25].

Misclassification in Relation to an Odds Ratio Example

This part studies potential influences of a misclassified binary exposure on OR. Assume the true data given in Table 1.1 are subject to misclassification: 5 percents of smokers are misclassified into the unexposed group, and 8% of nonsmoking individuals are miscategorized into the smoking category (Table 1.2). Note that this kind of misclassification belongs to the non-differential scenario because the same amount of misclassification is experienced by cases and controls. With the presence of misclassification, OR reduces to $\frac{396*245}{305*204}$ = 1.56 in contrast to 1.67 (without misclassification). This suggests OR is underestimated and this underestimation is referred to attenuation or reducing to the null[7].

Table 1.3: The effects of differential misclassification when 20% of smoking cases, but not controls, are misclassified as nonsmokers

	Lung Cancer	No Lung Cancer	Total
Smokers	400-80=320	300	620
Nonsmokers	200+80=280	250	530
Total	600	550	1150

Table 1.4: The effects of differential misclassification when 20% of nonsmoking cases, but not controls, are misclassified as smokers

	Lung Cancer	No Lung Cancer	Total
Smokers	400+40=440	300	740
Nonsmokers	200-40=160	250	410
Total	600	550	1150

Now suppose differential misclassification occurs in this case-control study. Assume that (1) 20% of the smokers who develop lung cancer are wrongly put into the nonsmoking group; (2) no nonsmoking cases are misclassified into the smoker category; and (3) no misclassification happens in the control group (Table 1.3). OR becomes 0.95 ($\frac{320*250}{300*280}$), which is toward to the null. In this instance, cases and controls as two groups receive different amounts of misclassification. More specifically, controls even do not have any misclassification. This situation may be due to the potential influence of the recall process (i.e. recall bias).

If the misclassification assumptions are altered so that 20 percents of the nonsmoking cases are misclassified into the exposed group and, again, no misclassification is within the controls (Table 1.4). Then OR turns to be 2.29, which indicates the parameter is overestimated or away from the null.

I have pointed out some methods in section 1.1 that can be used to reduce differential

(and non-differential) mismeasurement in the design stage. Generally speaking, the best way to avoid differential mismeasurement is to apply the same methodology and the same technology for every individual in the study[11].

1.3.3 Gold Standard and Validation

Gold standard is regarded as the best test available in the field that measures the true value of the exposure[26]. However, gold standard does not always exist due to the technical infeasibility. For instance, no gold standard is available to diagnose mental disorders in psychiatric studies[27]. In some circumstances, even the gold standard available, it is financially impractical to implement the best diagnostic test to every participant in the sample. Instead, researchers often first employ the cheap and simple but error-prone approach to the whole selected sample (N). And then apply the gold standard to a recruited small group (i.e. a validation sample). This validation sample can be selected either from the original sample with size n_1 (i.e. an internal validation) or from an external population with size n_2 (i.e. an external validation)[7]. In such a manner, we then have both true and proxy data for the validation group, which help correct misclassification in the main study group. The main study group is defined as the remaining individuals of the selected sample (N-n_1) in an internal validation study or the whole selected sample (N) from the original population in the context of an external validation[6]. External validation approach is primarily employed to generalize the results into other population. However, the main interest of observational studies is to make inferences about the original population where the main sample is chosen[19]. Therefore, the external validation is less useful than the internal validation for case-control and cohort studies. Note that, although this thesis does not deal with the gold standard and validation data, they are used a lot for real studies. Besides, the misclassification rates adopted in this thesis could be estimated from a validation group in practice.

The Nurses' Health Study described in section 1.1 employed the internal validation approach. Approximately 89,000 subjects were chosen as the main study group and exposures relating to nutritional intakes were measured through FFQ (i.e. an imprecise yet easy and cheap questionnaire). Later, an internal validation group (173) was selected and an gold

standard named as weighed diet records was implemented on every patient in the validation group. The weighed diet record measures nutritional intakes of various foods in a real-time basis regardless of the high cost it has. With the help of the validation data, researchers corrected misclassification in the main sample[16]. The *H.pylori* infection instance described in section 1.1.2 is another one employing the internal validation design.

Chapter 2

Background

The statistical literature with a focus of mismeasured independent variables has been extensively discussed over the last fifty years. Fuller[21] and Carroll et al.[7] provide systematic discussions on dealing with the mismeasurement problem. More specifically, Fuller[21] focuses on mismeasurement of explanatory variables in a linear regression model, and Carroll et al.[7] mainly study mismeasurement in a nonlinear regression context.

This chapter contains the following sections. Section 2.1 introduces the literature with a focus of consequences of ignoring misclassification. The remainder of the chapter concentrates on how to account for effects of binary misclassification (section 2.2) and multi-categorical misclassification (section 2.3), respectively.

2.1 Consequences of Ignoring Misclassification

Since Bross[28] published the first paper about the effects of non-differential misclassification in a binary variable (a correction by Newell[29]), many studies investigating the consequences of ignoring the misclassification have been done over the last fifty years. Some of the main findings resulted by naively ignoring misclassification are: (1) point estimates and standard errors can be biased[7]; (2) the actual association between two variables can be failed to detect[7]; and (3) the predictivity of the independent variable(s) on the response variable can be reduced[8].

In the following, I will explore the literature with more details. It has been proved by various studies that misclassification of a binary or multi-categorical exposure can bias the

estimate either toward or away from the null (one binary example given in section 1.3.2). It is also worth mentioning that, in earlier days, some researchers have wrongly stated that non-differential misclassification in a binary variable can only result in the attenuation of the estimate. However, the violations can be easily shown by the results of Davidov et al.[10] and this thesis as a special case. For instance, when the true odds ratio is smaller than one, misclassification results in an over-estimation under the non-differential assumption. On the other hand, although some of early studies provide a wrong conclusion, they can still be used to examine the consequences of ignoring binary misclassification to some extent. Copeland et al.[30] used instances to illustrate the changing trend of the estimated odds ratio or relative risk when they altered sensitivity and specificity, outcome prevalence and/or exposure occurrence for both non-differential and differential scenarios. Both under-estimated and overestimated situations were shown by their results. Greenland[31] analyzed the situation including two binary covariates and a perfectly measured binary response. The potential influences (i.e. under-estimation and over-estimation) due to misclassification of either one or two covariates were investigated by employed hypothetical numerical examples. Dosemecl et al.[32] used two hypothetical numerical instances to demonstrate, in the context of a misclassified multi-categorical exposure, the slope estimates could be underestimated or overestimated regardless of the misclassification scenario either non-differential or differential. Their conclusion for the non-differential situation was further proved by Birkett[33] using algebraic formulas. In addition, Weinberg et al.[34] explored conditions to keep the direction of the bias unchanged when non-differential misclassification took place in a multi-categorical exposure.

2.2 Correction Methods for Binary Misclassification

As a consequence of the fact that naively ignoring misclassification reduces the quality of inferences, a large amount of literature have been published for correcting misclassification. Two approaches are generally used, namely frequentist method and Bayesian method. Section 2.2.1 and 2.2.2 describe the statistical literature applying the frequentist approach to account for misclassification when a gold standard is either available or not. While Section 2.2.3

focuses on the literature employing the Bayesian methodology.

2.2.1 With Gold Standard

This subsection reviews some published works under the assumption of gold standard available. It is subdivided into two parts according to the criterion whether maximum likelihood (ML) method or other methods is used to estimate parameters.

Maximum Likelihood Method

Maximum likelihood method is widely used in the misclassification literature and in the real practice to obtain estimated parameters as this method is asymptotically unbiased and efficient under certain regularity conditions[35, 25, 6]. Besides, Prescott and Garthwaite[26] point out ML method gives higher power than other frequentist methods if the internal validation information is accessible.

Espeland and Hui[36] worked on maximum likelihood estimators(MLEs) and their variances when the gold standard was available and a log-linear model was built to relate a binary variable A to a binary variable B when either A or B or both was misclassified. Their method allowed the complex structure (e.g., the higher order interaction). Nevertheless, they admitted that "little rationale" for their instances used the log-linear model except simply following the previous published works[36]. Instead, some researchers pointed out the logistic regression model was employed much more often for the binary regression, which was a regression with a binary outcome[37, 6, 26]. Holcroft and Spiegelman[38] concentrated on choosing the optimal validation design in a sense to select the one with the minimum variance for MLE of the corrected odds ratio. Note that the log of odds ratio is equivalent to the coefficient of the exposure in a logistic regression (more in section 3.1). Under the assumption of the total size of the validation sample fixed, Holcroft and Spiegelman concluded it was preferable to adopt a balance design, which equally distributed validation individuals into four combinations of the outcome and the exposure, because of the "simplicity and good performance" the balance design achieved[38]. Spiegelman et al.[6] constructed a logistic regression model including covariates subject to measurement error, covariates prone to misclassification and

a list of error-free covariates. Their method requested to use the information from either internal or external validation data to estimate parameters by ML method. Spiegelman et al. also used the Nurses' Health Study and a simulation study to investigate both asymptotic and finite sample properties of their MLEs.

Other Frequentist Methods

Although MLEs gain asymptotic unbiasedness and efficiency under certain regularity conditions, it may be suggested not appropriate under some situations such as when regularity conditions violated, MLEs not exist, or sample size small enough so efficiency can not be achieved[35]. As a result, some other frequentist methods are also under investigation. Barron[39] employed the matrix method to estimate the true relative risk between two independent binary variables that are subject to non-differential misclassification. He assumed misclassification rates were known and fixed. Morrissey and Spiegelman[25] examined the differences of four methods on correcting misclassification for binary differential and non-differential scenarios when the gold standard was available to a validation sample. Four methods were Barron's[39] matrix method, inverse matrix method derived by Marshall[40], improved matrix method[40], and ML approach. The corresponding variances were worked out by Greenland[41] and Morrissey and Spiegelman[25]. The results showed the ML method was the most efficient one to correct misclassification. Lyles[42] later pointed out, in the context of differential binary misclassification, the inverse matrix method was equivalent to the ML method. Therefore, the inverse matrix method shared the same efficiency as the ML approach, but had a simpler computational process.

2.2.2 Without Gold Standard

Real world difficulties such as technical infeasibility and financial limitation could make a gold standard not available. The literature with a focus in such a situation is reviewed in this subsection. Walter and Irwig[14] did a review about the misclassification problem when a gold standard was not obtainable. They constructed various latent designs to ensure the

reliability of observed data. The latent class analysis refers to the analysis with the true value of the variable unknown. And that variable is named as latent variable[14]. Walter and Irwig stated, in order to make estimation free of constraints, at least three replicate observations per subject were needed.

Both Drews et al.[9] and Kosinski and Flanders[43] studied the case when there were two imperfect measures available for a binary exposure. Both of their studies calculated the corrected odds ratio through the expectation-maximization (EM) algorithm that helped obtain MLEs. Nonetheless, there were some key differences between two studies. First, Drews et al.'s research only dealt with the non-differential situation in a case-control study. While Kosinski and Flanders's work was able to dispose of both non-differential and differential misclassification in either a case-control study or a cohort study. Second, Kosinski and Flanders assumed the error between two imperfect tests was independent given the true status of the exposure. However, Drews et al. pointed out, if the independency was not actually true, the odds ratio could be underestimated. Therefore, Drews et al.'s work allowed the dependency of misclassification between the two imperfect tests when the degree of dependency was known. Third, Drews et al. did not include other error-free covariate(s), but Kosinski and Flanders did.

Lyes and Lin[44] calculated the range for estimated odds ratios or log of ORs by varying the value of sensitivity and specificity as a solution to a gold standard not available. Their method allowed both non-differential and differential misclassification. The corrected estimates in their paper were obtained by three methods: (1) matrix method by using sensitivity and specificity; (2) maximum likelihood approach; and (3) "inverse" matrix method by employing positive and negative predictive values (see section 1.3.2).

2.2.3 Bayesian Methods for Binary Misclassification

Bayesian approach is another primary method used to adjust the potential influences of misclassification on parameters. Different from frequentist method, Bayesian approach assumes

all parameters in the model are random. It requires the *priori* information about these parameters. It also asks for a likelihood of parameters, which can be expressed as the conditional distribution of data given parameters. The aim of Bayesian analysis is to obtain the posterior distribution with the help of the *priori* and the likelihood. The posterior distribution is the basis of all Bayesian inferences[45].

This subsection introduces the literature employing Bayesian methods to account for binary misclassification. The first part of this subsection studies the situation when misclassification is uncertain for researchers. And the second part focuses on the use of Bayesian methods for various study designs such as matched or unmatched case-control studies and cohort studies.

Misclassification Probabilities Unfixed

Under the frequentist context, the misclassification matrix is assumed fixed and known. However, in reality, we sometimes are uncertain about misclassification rates. Gustafson et al.[46] pointed out, if sensitivity and specificity had even small differences from actual probabilities, a large number of undetectable asymptotic biases relating to odds ratio could be induced in a case-control study. Therefore, Gustafson et al. suggested employing Bayesian method to incorporate the uncertainty of misclassification into the *priori* consideration. In their paper, Gustafson et al. constructed four independent priors for the prevalence of cases and controls, sensitivity, and specificity. Chu et al.[47] extended Gustafson et al.'s case by considering the correlation of sensitivity and specificity. They demonstrated how the posterior distribution altered along with the change of the *priori* under both non-differential and differential circumstances. Chu et al.[8] worked on the adjustment of one binary misclassified exposure in an unmatched case-control study when validation data are accessible. They compared the performance of their Bayesian estimators with those obtained through ML and SIMEX methods. They concluded Bayesian method had more merits in general than frequentist methods especially when the exposure was a rare event, the validation size was small, and misclassification of exposure whether differential or non-differential was not clear.

Bayesian Analysis for Various Study Design

Bayesian methods have been adopted to account for misclassification in exposure under various study designs such as cohort study, matched case-control study and unmatched case-control study. Prescott and Garthwaite investigated non-differential and differential misclassification of a binary covariate in unmatched case-control[48], prospective cohort[49], or matched case-control studies[26]. All their methods dealt with the case when validation data were available. For unmatched case-control studies, one Bayesian method with two stages was formulated to dispose of the situation where only one covariate is subject to misclassification. The first stage of their method analyzed the validation data; and the second focused on the main group data. Note that the *priori* of the second stage is the posterior of the first so that the information is transferred between these two stages. Prescott and Garthwaite also studied the misclassification issue in a prospective cohort study. The misclassified binary exposure and error-free continuous and categorical covariates were included in three logistic regression models, which were named as the disease model, the exposure model and the misclassification model. Weakly-informative priors that had large variances were chosen for all parameters in three models. The final results of their study allowed investigators to explore not only the relationship between the disease and the true exposure but also the association between the true and error-prone exposure and between the true exposure and all other covariates. Prescott and Garthwaite also published a paper highlighting the correction of misclassification in a matched case-control study. They examined the corrected OR by applying three models with different assumptions. The situation without validation data available in a matched case-control study was studied by Liu et al.[50] with the assumption of non-differential binary misclassification.

2.3 Correction Methods for Multiple-categorical Misclassification

This section introduces the literature that concentrates on the correction of misclassification in a multi-categorical exposure. Note that the asymptotic analysis for biases caused by mis-

classification in a multi-categorical variable has not been reported by the existing literature (at least to my best knowledge). Therefore, this thesis aims to fill this gap in the literature.

2.3.1 Frequentist Methods

K*ü*chenhoff et al.[51] proposed a misclassification simulation and extrapolation (MC-SIMEX) method to dispose of non-differential misclassification of a binary or multi-categorical exposure in a regression model. SIMEX method was originally designed for the measurement error problem, yet authors extended it to solve the misclassification problem when misclassification rates were fully knowledgeable or their estimates could be obtained from validation data. K*ü*chenhoff et al. also compared their method with the matrix method introduced by Morrissey and Spiegelman[25] when a binary misclassified exposure was in the model. Reade-Christopher and Kupper[52] investiaged the potential biases resulted by ignoring non-differential misclassification in a multi-categorical exposure in a follow-up study. Their method allowed other error-free covariates under consideration.

2.3.2 Bayesian Methods

There are relatively limited literature using Bayesian methods to handle multi-categorical misclassification. Kuroda and Zhi[53] concentrated on the situation where there were two multi-categorical variables A and B and B was subject to misclassification. The data augmentation (DA) algorithm was applied to find the posterior distribution and further to obtain the estimated Pr(A,B) (i.e. the joint probability of A and B), which was the aim of their study. Viana[54] studied misclassified multinomial data from a small sample and a two-stage Bayesian model was built. The first stage only analyzed the internal validation data; and the second one was for the data from the main study. Ruiz et al.[55] also proposed a Bayesian method to account for misclassification in a multinomial variable. They introduced latent vectors in the analysis in order to overcome the computational difficulty resulted by the likelihood and the posterior distribution.

Chapter 3

Asymptotic Bias

This chapter studies asymptotic biases caused by a misclassified multi-categorical exposure in a logistic regression context. It is organized as follows. Section 3.1 gives the details of deriving the bias formulas when only one misclassified multi-categorical exposure is in the model. Section 3.2 figures out the bias formulas when there is one more perfectly measured binary covariate in the model.

3.1 A Single Misclassified Exposure in the Model

3.1.1 Notations and Models

Assume E is a multi-categorical exposure with m categories. Its distribution is specified as $P(E = j) = p_j,\ j = 0, 1, ..., m-1$; with the constraint $\sum_{j=0}^{m-1} p_j = 1$. Define a list of indicator variables for E; that is,

$$E_t = \begin{cases} 1, & \text{if } E = t,\ t = 1, 2, ..., m-1; \\ 0, & \text{otherwise.} \end{cases}$$

Suppose that a logistic regression model is built to relate the error-free binary outcome Y to the predictors (i.e. E_t) as follows:

$$\text{logit}P(Y = 1|E) = \beta_0 + \beta_1 E_1 + \beta_2 E_2 + \cdots + \beta_{m-1} E_{m-1}. \tag{3.1}$$

Note that the regression coefficients β_0, β_1, ...,β_{m-1} have some potential meanings relating to the risk of developing a certain disease. The details are presented in section 3.1.2.

As I pointed out in Chapter 1, it is common that, in practice, various reasons could lead the exposure surrogate to misclassification. If this is the case, we then have the proxy value of E, denoted by X. Similarly, X can be expressed in terms of a set of indicator variables:

$$X_t = \begin{cases} 1, & \text{if } X = t,\ t = 1, 2, ..., m-1; \\ 0, & \text{otherwise.} \end{cases}$$

Hence the logistic model for (X, Y) is:

$$\text{logit}P(Y = 1|X) = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \cdots + \gamma_{m-1} X_{m-1}. \tag{3.2}$$

Let $\theta_{d,ij}$ represent the probability that an individual from category i $(i = 0, 1, ...m-1)$ classified into category j given the outcome d $(d = 0, 1)$. That is,

$$\theta_{d,ij} \quad = \quad P(X = i|E = j, Y = d).$$

As described in Chapter 1, $\boldsymbol{\theta}$ represent misclassification rates and they can be re-written into a matrix form:

$$\theta \quad = \quad \begin{pmatrix} \theta_{d,00} & \theta_{d,01} & \cdots & \theta_{d,0(m-1)} \\ \theta_{d,10} & \theta_{d,11} & \cdots & \theta_{d,1(m-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{d,(m-1)0} & \theta_{d,(m-1)1} & \cdots & \theta_{d,(m-1)(m-1)} \end{pmatrix}.$$

Each diagonal value of $\boldsymbol{\theta}$ represents how likely we have the exposure recorded correctly; and each off-diagonal value indicates the misclassification probability. Therefore, the higher the diagonal values are, the less severe misclassification problem we encounter with. Note that the sum of values in each column of the misclassification matrix is one by the definition of $\theta_{d,ij}$. For the binary case that Davidov et al.[10] have studied, the above misclassification matrix can be reduced to a 2×2 matrix with two parameters, namely sensitivity $P(X = 1|E = 1, Y = d)$ and specificity $P(X = 0|E = 0, Y = d)$.

Here are some notations that will be used in the bias evaluation in section 3.1.2 and 3.1.3.

Define:

$$\pi_{dj} = P(Y = d, E = j),$$
$$p_{di} = P(Y = d, X = i),$$
$$\pi_{d|j} = P(Y = d|E = j),$$
$$p_{d|i} = P(Y = d|X = i),$$
$$\xi_{d,ab} = \frac{\pi_{da}}{\pi_{db}} = \frac{P(E = a|Y = d)}{P(E = b|Y = d)},$$

where a, b=0,1,..,m-1.

By applying the definition of conditional probability, we have:

$$\begin{aligned} p_{di} &= \sum_{j=0}^{m-1} P(X = i, E = j, Y = d) \\ &= \sum_{j} P(X = i|E = j, Y = d)P(Y = d, E = j) \\ &= \sum_{j} \theta_{d,ij}\pi_{dj}. \end{aligned} \tag{3.3}$$

This relationship can be re-arranged into a matrix form:

$$\mathbf{p}_d = \boldsymbol{\theta}_d \boldsymbol{\pi}_d,$$

where $\mathbf{p}_d = (p_{d0}, p_{d1}, \ldots, p_{d(m-1)})', \boldsymbol{\theta}_d = (\theta_{d,ij}), \boldsymbol{\pi}_d = (\pi_{d0}, \pi_{d1}, \ldots, \pi_{d(m-1)})'$.
In such a way, equation (3.3) builds a connection between $\boldsymbol{\pi}_d$ and $\mathbf{p}_d$ via $\boldsymbol{\theta}_d$.

In addition, by the definition of conditional probability, we have:

$$\frac{\pi_{1|j}}{\pi_{0|j}} = \frac{\pi_{1j}}{\pi_{0j}} \quad \text{and} \quad \frac{p_{1|j}}{p_{0|j}} = \frac{p_{1j}}{p_{0j}}.$$

3.1.2 Asymptotic Bias under Differential Misclassification

In this subsection, I derive formulas for asymptotic biases due to differential misclassification in a multi-categorical exposure. First, I present the regression coefficients in the following:

$$\begin{aligned} \beta_0 &= \text{logit}P(Y = 1|E = 0), \\ \beta_0 + \beta_t &= \text{logit}P(Y = 1|E = t), \end{aligned}$$

$$
\begin{aligned}
e^{\beta_0} &= \text{odds}\big(P(Y=1|E=0)\big) \\
&= \frac{P(Y=1|E=0)}{P(Y=0|E=0)},
\end{aligned}
$$

$$
\begin{aligned}
e^{\beta_t} &= \frac{\text{odds}\big(P(Y=1|E=t)\big)}{\text{odds}\big(P(Y=1|E=0)\big)} \\
&= \frac{P(Y=1|E=t)}{P(Y=0|E=t)} \frac{P(Y=0|E=0)}{P(Y=1|E=0)} \\
&= \frac{P(E=t|Y=1)}{P(E=0|Y=1)} \frac{P(E=0|Y=0)}{P(E=t|Y=0)} \\
&= \frac{\pi_{1t}}{\pi_{10}} \frac{\pi_{00}}{\pi_{0t}} \\
&= \frac{\xi_{1,t0}}{\xi_{0,t0}}.
\end{aligned}
$$

Analogously, we have the expression for the γ_i's in model (3.2):

$$
\begin{aligned}
\gamma_0 &= \text{logit}P(Y=1|X=0), \\
\gamma_0+\gamma_t &= \text{logit}P(Y=1|X=t).
\end{aligned}
$$

Let Δ_i denote the asymptotic bias, which equals to $\beta_i - \gamma_i$. We then have:

$$
\begin{aligned}
\Delta_0 &= \beta_0 - \gamma_0 \\
&= \text{logit}(\pi_{1|0}) - \text{logit}(p_{1|0}) \\
&= \log\big(\frac{\pi_{1|0}}{\pi_{0|0}}\big) - \log\big(\frac{p_{1|0}}{p_{0|0}}\big) \\
&= \log\big(\frac{\pi_{10}}{\pi_{00}}\big) - \log\big(\frac{p_{10}}{p_{00}}\big) \\
&= \log\left(\frac{\pi_{10}}{\pi_{00}} \frac{\sum_{j=0}^{m-1}\theta_{0,0j}\pi_{0j}}{\sum_{j=0}^{m-1}\theta_{1,0j}\pi_{1j}}\right) \\
&= \log\left(\frac{\sum_{j=0}^{m-1}\theta_{0,0j}\xi_{0,j0}}{\sum_{j=0}^{m-1}\theta_{1,0j}\xi_{1,j0}}\right). \qquad (3.4)
\end{aligned}
$$

Similarly, Δ_t can be expressed as a function of misclassification rates:

$$\begin{aligned}
\Delta_t &= \beta_t - \gamma_t \\
&= \text{logit}(\pi_{1|t}) - \text{logit}(p_{1|t}) - (\beta_0 - \gamma_0) \\
&= \log\Big(\frac{\pi_{1t}p_{0t}}{\pi_{0t}p_{1t}}\Big) - \Delta_0 \\
&= \log\left(\frac{\pi_{1t}}{\pi_{0t}}\frac{\sum_j \theta_{0,tj}\pi_{0j}}{\sum_j \theta_{1,tj}\pi_{1j}}\right) - \Delta_0 \\
&= \log\left(\frac{\sum_{j=0}^{m-1} \theta_{0,tj}\xi_{0,jt}}{\sum_{j=0}^{m-1} \theta_{1,tj}\xi_{1,jt}}\right) - \Delta_0. \qquad (3.5)
\end{aligned}$$

In the same spirit of Davidov et al.[10], I simplify the formulas (3.4) and (3.5) by doing the first order Taylor expansion to them at $\boldsymbol{\theta}_d = \mathbf{I}_m$, where $\boldsymbol{\theta}_d = \mathbf{I}_m$ stands for no misclassification. Thus, the approximated formulas are:

$$\Delta_0 \approx \sum_{j=0}^{m-1}(\theta_{0,0j}\xi_{0,j0} - \theta_{1,0j}\xi_{1,j0}), \qquad (3.6)$$

$$\Delta_t \approx \sum_{j=0}^{m-1}\big(\theta_{0,tj}\xi_{0,jt} - \theta_{1,tj}\xi_{1,jt}\big) - \sum_{j=0}^{m-1}(\theta_{0,0j}\xi_{0,j0} - \theta_{1,0j}\xi_{1,j0}). \qquad (3.7)$$

3.1.3 Asymptotic Bias under Non-differential Misclassification

This subsection focuses on the scenario where misclassification in exposure does not provide additional information about Y, which refers to non-differential misclassification as stated in section 1.2.3. Since non-differential misclassification indicates the misclassification in exposure does not depend on the disease status, we then have $\theta_{0,ij} = \theta_{1,ij} = \theta_{ij}$. Thus, the bias derivation can be simplified as:

$$\begin{aligned}
\Delta_0 &= \log\left(\frac{\sum_{j=0}^{m-1} \theta_{0j}\xi_{0,j0}}{\sum_{j=0}^{m-1} \theta_{0j}\xi_{1,j0}}\right) \\
&= \log\left(\frac{\sum_{j=0}^{m-1} \theta_{0j}\xi_{0,j0}}{\sum_{j=0}^{m-1} \theta_{0j}\frac{\xi_{1,j0}}{\xi_{0,j0}}\xi_{0,j0}}\right) \\
&= \log\left(\frac{\theta_{00} + \sum_{j=1}^{m-1} \theta_{0j}\xi_{0,j0}}{\theta_{00} + \sum_{j=1}^{m-1} \theta_{0j}\exp(\beta_j)\xi_{0,j0}}\right). \qquad (3.8)
\end{aligned}$$

Note that there is no asymptotic bias for the intercept (i.e. $\Delta_0 = 0$) if $\beta_j = 0$ $(j \neq 0)$. To put it in another way, no bias is obtained if the odds ratio of $Y = 1$ between category j and category

0 is one. Furthermore, this also implies the exposure E and the response Y are independent each other. Consequently, E should not be added into the model.

In practice, researchers may be interested in the direction of the bias. That is, whether misclassification results in over-estimation or under-estimation. According to equation (3.8), we have $\Delta_0 > 0$ when $\beta_j < 0$, and $\Delta_0 < 0$ when $\beta_j > 0$. This is to say, the estimated regression coefficient $\widehat{\gamma_0}$ is (asymptotically) attenuated when β_j are negative or when odds ratios at $Y = 1$ between category j and category 0 are smaller than one. On the other hand, $\widehat{\gamma_0}$ is (asymptotically) inflated when β_j are positives or when the odds of Y=1 for category j are greater than the odds for category 0. It is worth to note that the conditions described above on β_j are sufficient but not necessary. That is, for example, it is not necessary to have all negative β_j to ensure a positive Δ_0[56].

Similarly,

$$\begin{aligned}
\Delta_t &= \log\left(\frac{\sum_{j=0}^{m-1}\theta_{tj}\xi_{0,jt}}{\sum_{j=0}^{m-1}\theta_{tj}\xi_{1,jt}}\right) - \log\left(\frac{\sum_{j=0}^{m-1}\theta_{0j}\xi_{0,j0}}{\sum_{j=0}^{m-1}\theta_{0j}\xi_{1,j0}}\right) \\
&= \log\left(\frac{\theta_{tt} + \sum_{j\neq t}\theta_{tj}\xi_{0,jt}}{\theta_{tt} + \sum_{j\neq t}\theta_{tj}\frac{\xi_{1,jt}}{\xi_{0,jt}}\xi_{0,jt}}\right) - \Delta_0 \\
&= \log\left(\frac{\theta_{tt} + \sum_{j\neq t}\theta_{tj}\xi_{0,jt}}{\theta_{tt} + \sum_{j\neq t}\theta_{tj}\exp(\beta_j^* - \beta_t)\xi_{0,jt}}\right) - \\
&\quad \log\left(\frac{\theta_{00} + \sum_{j=1}^{m-1}\theta_{0j}\xi_{0,j0}}{\theta_{00} + \sum_{j=1}^{m-1}\theta_{0j}\exp(\beta_j)\xi_{0,j0}}\right), \qquad (3.9)
\end{aligned}$$

where $\beta_j^* = \beta_j$ if $j > 0$ and is zero otherwise.

Note that $\Delta_t = 0$ when all β_j $(j \neq 0)$ equal to zero, which is the same sufficient condition to ensure $\Delta_0 = 0$.

Let $\boldsymbol{\beta}_{-t} = \{\beta_1, \beta_2, \ldots, \beta_{t-1}, \beta_{t+1}, \ldots, \beta_{m-1}\}$, i.e. the collection of all regression coefficients except β_0 and β_t. If $0 < \boldsymbol{\beta}_{-t} < \beta_t$ then the first term of Δ_t is positive and Δ_0 is negative. Therefore, it is a sufficient condition to ensure a positive Δ_t. On the other hand, if $\beta_t < \boldsymbol{\beta}_{-t} < 0$ then the first term of Δ_t is negative and Δ_0 is positive, which together result in a negative

Δ_t. In other words, if all odds of Y=1 for category j are larger than the odds for category 0 and the odds for category t has the largest odds, then we have attenuated $\widehat{\gamma_t}$. While we have inflated $\widehat{\gamma_t}$ when the odds of Y=1 for category 0 is the biggest odds and the odds for category t is the smallest odds[56].

Similar to Davidov et al.'s[10] work toward the simplification of the bias expressions, I apply Taylor approximation to formulas (3.8) and (3.9) around $\boldsymbol{\theta}_d = \mathbf{I}_m$ and have:

$$\Delta_0 \approx \sum_{j=0}^{m-1} (\theta_{0j}\xi_{0,j0} - \theta_{0j}\xi_{1,j0}), \tag{3.10}$$

$$\Delta_t \approx \sum_{j=0}^{m-1} \left(\theta_{tj}\xi_{0,jt} - \theta_{tj}\xi_{1,jt}\right) - \sum_{j=0}^{m-1} (\theta_{0j}\xi_{0,j0} - \theta_{0j}\xi_{1,j0}). \tag{3.11}$$

3.2 Perfectly Measured Covariate Included in the Model

3.2.1 Notations and Models

In this part, I study a different situation where both a multi-categorical variable, E, and an error-free binary covariate, Z, are included in the model. A logistic regression is built between the outcome Y and the covariatess Z and E_t as follows:

$$\text{logit}P(Y = 1|E, Z) = \beta_0 + \beta_1 E_1 + \beta_2 E_2 + \cdots + \beta_{m-1} E_{m-1} + \beta_m Z. \tag{3.12}$$

Note no interaction is assumed between Z and E_t.

Analogously, a logistic regression model is built for (X, Z, Y):

$$\text{logit}P(Y = 1|X, Z) = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \cdots + \gamma_{m-1} X_{m-1} + \gamma_m Z. \tag{3.13}$$

Similar to the notations introduced in section 3.1.1, I define the following.

For k=0,1;

$$\pi_{dj|k} = P(Y = d, E = j|Z = k),$$
$$p_{di|k} = P(Y = d, X = i|Z = k),$$
$$\pi_{d|jk} = P(Y = d|E = j, Z = k),$$
$$p_{d|ik} = P(Y = d|X = i, Z = k),$$
$$\xi_{d,ab|k} = \frac{\pi_{da|k}}{\pi_{db|k}} = \frac{P(E = a, Y = d|Z = k)}{P(E = b, Y = d|Z = k)} = \frac{P(E = a|Y = d, Z = k)}{P(E = b|Y = d, Z = k)}.$$

Assume X and Z are independent conditional on E and Y, i.e., $P(X = i|E = j, Y = d, Z = k) = P(X = i|E = j, Y = d) = \theta_{d,ij}$. This assumption indicates misclassification is free of Z. Therefore, misclassification rates have the same form and the same explanation as section 3.1.1 described.

By the definition of conditional probability, we also have:

$$\begin{aligned} p_{di|k} &= \sum_{j=0}^{m-1} P(X = i, E = j, Y = d|Z = k) \\ &= \sum_{j} P(X = i|E = j, Y = d, Z = k)P(Y = d, E = j|Z = k) \\ &= \sum_{j} P(X = i|E = j, Y = d)\pi_{dj|k} \\ &= \sum_{j} \theta_{d,ij}\pi_{dj|k}. \end{aligned} \tag{3.14}$$

This relationship can be re-arranged into a matrix form:

$$\mathbf{p}_{d|k} = \boldsymbol{\theta}_d \boldsymbol{\pi}_{d|k},$$

where $\mathbf{p}_{d|k} = (p_{d0|k}, p_{d1|k}, \ldots, p_{d(m-1)|k})'$, $\boldsymbol{\theta}_d = (\theta_{d,ij})$, $\boldsymbol{\pi}_{d|k} = (\pi_{d0|k}, \pi_{d1|k}, \ldots, \pi_{d(m-1)|k})'$.

Moreover, by the definition of conditional probability, we have:

$$\frac{\pi_{1|jk}}{\pi_{0|jk}} = \frac{\pi_{1j|k}}{\pi_{0j|k}} \text{ and } \frac{p_{1|jk}}{p_{0|jk}} = \frac{p_{1j|k}}{p_{0j|k}}.$$

3.2.2 Asymptotic Bias under Differential Misclassification

This subsection gives the expressions for asymptotic biases (i.e. $\Delta_c = \beta_c - \gamma_c$, $c = 0, 1, ..., m$) when an error-free covariate, denoted by Z, is involved in the study.

The regression coefficients in model (3.12) can be re-written as:

$$\begin{aligned}
\beta_0 &= \text{logit}P(Y = 1|E = 0, Z = 0), \\
\beta_0 + \beta_t &= \text{logit}P(Y = 1|E = t, Z = 0), \\
\beta_0 + \beta_m &= \text{logit}P(Y = 1|E = 0, Z = 1).
\end{aligned}$$

Thus, the exponential of β_0, β_t, and β_m can be expressed as below:

$$\begin{aligned}
e^{\beta_0} &= \text{odds}\big(P(Y = 1|E = 0, Z = 0)\big) \\
&= \frac{P(Y = 1|E = 0, Z = 0)}{P(Y = 0|E = 0, Z = 0)}; \\
e^{\beta_t} &= \frac{\text{odds}\big(P(Y = 1|E = t, Z = 0)\big)}{\text{odds}\big(P(Y = 1|E = 0, Z = 0)\big)} \\
&= \frac{P(Y = 1|E = t, Z = 0)}{P(Y = 0|E = t, Z = 0)} \frac{P(Y = 0|E = 0, Z = 0)}{P(Y = 1|E = 0, Z = 0)} \\
&= \frac{P(E = t|Y = 1, Z = 0)}{P(E = 0|Y = 1, Z = 0)} \frac{P(E = 0|Y = 0, Z = 0)}{P(E = t|Y = 0, Z = 0)} \\
&= \frac{\xi_{1,t0|0}}{\xi_{0,t0|0}}, \\
e^{\beta_m} &= \frac{\text{odds}\big(P(Y = 1|E = 0, Z = 1)\big)}{\text{odds}\big(P(Y = 1|E = 0, Z = 0)\big)} \\
&= \frac{P(Y = 1|E = 0, Z = 1)}{P(Y = 0|E = 0, Z = 1)} \frac{P(Y = 0|E = 0, Z = 0)}{P(Y = 1|E = 0, Z = 0)} \\
&= \frac{P(Y = 1, E = 0|Z = 1)}{P(Y = 0, E = 0|Z = 1)} \frac{P(Y = 0, E = 0|Z = 0)}{P(Y = 1, E = 0|Z = 0)}.
\end{aligned}$$

In the same manner, the regression coefficients γ_c in model (3.13) can be written as:

$$\begin{aligned}
\gamma_0 &= \text{logit}P(Y = 1|X = 0, Z = 0), \\
\gamma_0 + \gamma_t &= \text{logit}P(Y = 1|X = t, Z = 0), \\
\gamma_0 + \gamma_m &= \text{logit}P(Y = 1|X = 0, Z = 1).
\end{aligned}$$

With the above information, I derive the formula for the bias in intercept as follows:

$$\begin{aligned}
\Delta_0 &= \beta_0 - \gamma_0 \\
&= \text{logit}(\pi_{1|00}) - \text{logit}(p_{1|00}) \\
&= \log\Big(\frac{\pi_{1|00}}{\pi_{0|00}}\Big) - \log\Big(\frac{p_{1|00}}{p_{0|00}}\Big) \\
&= \log\Big(\frac{\pi_{10|0}}{\pi_{00|0}}\Big) - \log\Big(\frac{p_{10|0}}{p_{00|0}}\Big) \\
&= \log\left(\frac{\pi_{10|0}}{\pi_{00|0}}\frac{\sum_j \theta_{0,0j}\pi_{0j|0}}{\sum_j \theta_{1,0j}\pi_{1j|0}}\right) \\
&= \log\left(\frac{\sum_{j=0}^{m-1} \theta_{0,0j}\xi_{0,j0|0}}{\sum_{j=0}^{m-1} \theta_{1,0j}\xi_{1,j0|0}}\right). \qquad (3.15)
\end{aligned}$$

Similarly, I obtain the biases for the regression coefficients of E_t:

$$\begin{aligned}
\Delta_t &= \beta_t - \gamma_t \\
&= \text{logit}(\pi_{1|t0}) - \text{logit}(p_{1|t0}) - (\beta_0 - \gamma_0) \\
&= \log\Big(\frac{\pi_{1t|0}p_{0t|0}}{\pi_{0t|0}p_{1t|0}}\Big) - \Delta_0 \\
&= \log\left(\frac{\pi_{1t|0}}{\pi_{0t|0}}\frac{\sum_j \theta_{0,tj}\pi_{0j|0}}{\sum_j \theta_{1,tj}\pi_{1j|0}}\right) - \Delta_0 \\
&= \log\left(\frac{\sum_{j=0}^{m-1} \theta_{0,tj}\xi_{0,jt|0}}{\sum_{j=0}^{m-1} \theta_{1,tj}\xi_{1,jt|0}}\right) - \Delta_0. \qquad (3.16)
\end{aligned}$$

The asymptotic bias for the regression coefficient of the error-free Z is:

$$\begin{aligned}
\Delta_m &= \beta_m - \gamma_m \\
&= \text{logit}(\pi_{1|01}) - \text{logit}(p_{1|01}) - (\beta_0 - \gamma_0) \\
&= \log\Big(\frac{\pi_{10|1}p_{00|1}}{\pi_{00|1}p_{10|1}}\Big) - \Delta_0 \\
&= \log\left(\frac{\pi_{10|1}}{\pi_{00|1}}\frac{\sum_j \theta_{0,0j}\pi_{0j|1}}{\sum_j \theta_{1,0j}\pi_{1j|1}}\right) - \Delta_0 \\
&= \log\left(\frac{\sum_{j=0}^{m-1} \theta_{0,0j}\xi_{0,j0|1}}{\sum_{j=0}^{m-1} \theta_{1,0j}\xi_{1,j0|1}}\right) - \Delta_0. \qquad (3.17)
\end{aligned}$$

The expressions (3.15)-(3.17) can be further simplified by taking the first order Taylor ex-

pansion about $\boldsymbol{\theta}_d = \mathbf{I}_m$. Thus, the approximate bias formulas are obtained as below:

$$\Delta_0 \approx \sum_{j=0}^{m-1} (\theta_{0,0j}\xi_{0,j0|0} - \theta_{1,0j}\xi_{1,j0|0}), \tag{3.18}$$

$$\Delta_t \approx \sum_{j=0}^{m-1} \left(\theta_{0,tj}\xi_{0,jt|0} - \theta_{1,tj}\xi_{1,jt|0}\right) - \sum_{j=0}^{m-1} (\theta_{0,0j}\xi_{0,j0|0} - \theta_{1,0j}\xi_{1,j0|0}), \tag{3.19}$$

$$\Delta_m \approx \sum_{j=0}^{m-1} \left(\theta_{0,0j}\xi_{0,j0|1} - \theta_{1,0j}\xi_{1,j0|1}\right) - \sum_{j=0}^{m-1} (\theta_{0,0j}\xi_{0,j0|0} - \theta_{1,0j}\xi_{1,j0|0}). \tag{3.20}$$

3.2.3 Asymptotic Bias under Non-differential Misclassification

This subsection evaluates the asymptotic biases when a multi-categorical exposure variable, E, in model (3.12) is subject to non-differential misclassification. The non-differential scenario indicates $\theta_{0,ij} = \theta_{1,ij} = \theta_{ij}$. Therefore, the bias in intercept is simplified as below:

$$\begin{aligned}
\Delta_0 &= \log\left(\frac{\sum_{j=0}^{m-1}\theta_{0j}\xi_{0,j0|0}}{\sum_{j=0}^{m-1}\theta_{0j}\xi_{1,j0|0}}\right) \\
&= \log\left(\frac{\sum_{j=0}^{m-1}\theta_{0j}\xi_{0,j0|0}}{\sum_{j=0}^{m-1}\theta_{0j}\frac{\xi_{1,j0|0}}{\xi_{0,j0|0}}\xi_{0,j0|0}}\right) \\
&= \log\left(\frac{\theta_{00} + \sum_{j=1}^{m-1}\theta_{0j}\xi_{0,j0|0}}{\theta_{00} + \sum_{j=1}^{m-1}\theta_{0j}\exp(\beta_j)\xi_{0,j0|0}}\right).
\end{aligned} \tag{3.21}$$

Equation (3.21) implies there is no bias for the intercept if $\beta_j = 0$ $(j \neq 0)$. It also implies that Δ_0 is zero when E is not related to Y.

I also investigate the directions of the biases. Based on equation (3.21), we have $\Delta_0 < 0$ when $\beta_j > 0$ and $\Delta_0 > 0$ when $\beta_j < 0$. Equivalently, for all $j \neq 0$, if the odds of Y=1 and Z=0 for category j is bigger than the odds for category 0, then $\widehat{\gamma_0}$ is (asymptotically) inflated. On the other hand, if the odds of Y=1 and Z=0 for category j is smaller than the odds for category 0, then $\widehat{\gamma_0}$ is (asymptotically) attenuated. Note that the conditions that determine the signs of Δ_0 are not necessary but sufficient.

Under non-differential misclassification, Δ_t can be written as follows:

$$\begin{aligned}
\Delta_t &= \log\left(\frac{\sum_{j=0}^{m-1}\theta_{tj}\xi_{0,jt|0}}{\sum_{j=0}^{m-1}\theta_{tj}\xi_{1,jt|0}}\right) - \Delta_0 \\
&= \log\left(\frac{\theta_{tt} + \sum_{j\neq t}\theta_{tj}\xi_{0,jt|0}}{\theta_{tt} + \sum_{j\neq t}\theta_{tj}\frac{\xi_{1,jt|0}}{\xi_{0,jt|0}}\xi_{0,jt|0}}\right) - \Delta_0 \\
&= \log\left(\frac{\theta_{tt} + \sum_{j\neq t}\theta_{tj}\xi_{0,jt|0}}{\theta_{tt} + \sum_{j\neq t}\theta_{tj}\exp(\beta_j^* - \beta_t)\xi_{0,jt|0}}\right) - \\
&\quad \log\left(\frac{\theta_{00} + \sum_{j=1}^{m-1}\theta_{0j}\xi_{0,j0|0}}{\theta_{00} + \sum_{j=1}^{m-1}\theta_{0j}\exp(\beta_j)\xi_{0,j0|0}}\right),
\end{aligned} \tag{3.22}$$

where $\beta_j^* = \beta_j$ if $j \in (1, 2, ..., m-1)$ and is zero otherwise.

Same as the Δ_0 case, we have the unbiasedness when $\beta_j = 0$ $(j \neq 0)$ or equivalently, when the corresponding odds ratios equal to one.

Let us define $\boldsymbol{\beta}_{-t} = \{\beta_1, \beta_2, \ldots, \beta_{t-1}, \beta_{t+1}, \ldots, \beta_{m-1}\}$, i.e. the collection of all regression coefficients except β_0, β_t and β_m. According to expression (3.22), the first term of Δ_t is positive if all $\boldsymbol{\beta}_{-t}$ are smaller than β_t and $\boldsymbol{\beta}_{-t} > 0$. The first term of Δ_t is negative if all $\boldsymbol{\beta}_{-t}$ are larger than β_t and $\boldsymbol{\beta}_{-t} < 0$. Note that these conditions are sufficient to keep the signs of Δ_t as well. That is, $0 < \boldsymbol{\beta}_{-t} < \beta_t$ is sufficient to have a positive Δ_t and $\beta_t < \boldsymbol{\beta}_{-t} < 0$ is sufficient to have a negative Δ_t. Alternatively, we can say, if the odds of Y=1 and Z=0 for category t has the largest odds and the odds of Y=1 and Z=0 for category 0 has the smallest odds, we have attenuated $\widehat{\gamma_t}$. And, if the odds of Y=1 and Z=0 for category 0 is the biggest odds and the odds for category t is the smallest odds, we have inflated $\widehat{\gamma_t}$.

The asymptotic bias in the coefficient of Z is:

$$\Delta_m = \log\left(\frac{\sum_{j=0}^{m-1}\theta_{0j}\xi_{0,j0|1}}{\sum_{j=0}^{m-1}\theta_{0j}\xi_{1,j0|1}}\right) - \log\left(\frac{\sum_{j=0}^{m-1}\theta_{0j}\xi_{0,j0|0}}{\sum_{j=0}^{m-1}\theta_{0j}\xi_{1,j0|0}}\right). \tag{3.23}$$

The first term of Δ_m equals zero if $\xi_{0,j0|1} = \xi_{1,j0|1}$. In addition, $\xi_{0,j0|1} = \xi_{0,j0|0}$ and $\xi_{1,j0|1} = \xi_{1,j0|0}$ are sufficient conditions to have $\Delta_m = 0$. Therefore, the assumption Z is independent of E and Y ensures that misclassification of E has no effects on the coefficient of Z. Yet this condition also suggests Z should not be included in the model defined in equation (3.12).

Moreover, for all $j \in (1, ..., m-1)$, $\xi_{0,j0|1} > \xi_{1,j0|1}$ indicates the first term of Δ_m is positive and $\xi_{0,j0|1} < \xi_{1,j0|1}$ implies the first term of Δ_m is negative.

Under the non-differential misclassification assumption, the first order approximations are:

$$\Delta_0 \approx \sum_{j=0}^{m-1} (\theta_{0j}\xi_{0,j0|0} - \theta_{0j}\xi_{1,j0|0}), \tag{3.24}$$

$$\Delta_t \approx \sum_{j=0}^{m-1} \left(\theta_{tj}\xi_{0,jt|0} - \theta_{tj}\xi_{1,jt|0}\right) - \sum_{j=0}^{m-1} (\theta_{0j}\xi_{0,j0|0} - \theta_{0j}\xi_{1,j0|0}), \tag{3.25}$$

$$\Delta_m \approx \sum_{j=0}^{m-1} \left(\theta_{0j}\xi_{0,j0|1} - \theta_{0j}\xi_{1,j0|1}\right) - \sum_{j=0}^{m-1} (\theta_{0j}\xi_{0,j0|0} - \theta_{0j}\xi_{1,j0|0}). \tag{3.26}$$

Chapter 4

Examples and Simulation Studies

To better understand the asymptotic results obtained in Chapter 3, I employ illustrative examples and also conduct simulation studies in this chapter. The examples help understand how the asymptotic biases for regression coefficients alter along with the change of misclassification parameters. While the simulation studies help verify the validity of the asymptotic results by comparing the estimated biases from simulated datasets and the asymptotic values from the derived formulas.

4.1 Illustrative Examples

This section provides examples to illustrate how asymptotic biases can be influenced by misclassification rates. A three-categorical exposure variable is used for the demonstration.

4.1.1 Only One Misclassified Exposure in the Model

The focus of this subsection is to observe the asymptotic biases. At the beginning, I build a logistic regression model to relate the true exposure E to the response Y:

$$\text{logit}P(Y = 1|E) = \beta_0 + \beta_1 E_1 + \beta_2 E_2. \tag{4.1}$$

Similarly, a model is built to relate the proxy X to Y:

$$\text{logit}P(Y = 1|X) = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2. \tag{4.2}$$

Assume the marginal distribution of E is $P(E = 0) = 0.4$, $P(E = 1) = 0.2$, and $P(E = 2) = 0.4$.

Define a list of hypothetical misclassification matrices. Suppose that misclassification only occurs in the adjacent category or adjacent categories with equal probability for each category. That is, for each d $\in$ (0,1), if $P(X = 0|E = 0, Y = d) = a$, then $P(X = 1|E = 0, Y = d) = 1 - a$; if $P(X = 1|E = 1, Y = d) = b$, then $P(X = 0|E = 1, Y = d) = P(X = 2|E = 1, Y = d) = (1 - b)/2$; and if $P(X = 2|E = 2, Y = d) = c$, then $P(X = 1|E = 2, Y = d) = 1 - c$. Therefore, if we have $\theta_{d,ii} = (a, b, c)$, then the corresponding misclassification matrix $\boldsymbol{\theta}$ for $Y = d$ is:

$$\begin{pmatrix} a & (1-b)/2 & 0 \\ 1-a & b & 1-c \\ 0 & (1-b)/2 & c \end{pmatrix}.$$

Based on the above definition of $\theta_{d,ii}$ and the misclassification matrix $\boldsymbol{\theta}$, three pairs of misclassification matrices (i.e. Type 1, Type 2, Type 3) are employed for the differential scenario in this chapter. As I introduced in section 1.3.2, the diagonal values of a misclassification matrix (i.e. $\theta_{d,ii}$) represent the correctly distributed probabilities (i.e. classification probabilities) and the off-diagonal ones indicate the misclassification probabilities. Therefore, from Type 1 to Type 3, we can see the number of errors are assigned in an ascending order:

Type 1: $\theta_{0,ii} = (0.9, 0.9, 0.9)$ and $\theta_{1,ii} = (0.8, 0.8, 0.8)$;

Type 2: $\theta_{0,ii} = (0.9, 0.65, 0.9)$ and $\theta_{1,ii} = (0.7, 0.7, 0.8)$;

Type 3: $\theta_{0,ii} = (0.9, 0.65, 0.65)$ and $\theta_{1,ii} = (0.7, 0.7, 0.7)$.

Six types of misclassification are considered for the non-differential situation. Three of them are organized similarly to the differential case while the number of errors are arranged in an ascending order:

Type 4: $\theta_{ii} = (0.9, 0.9, 0.9)$,

Type 5: $\theta_{ii} = (0.9, 0.65, 0.9)$,

Type 6: $\theta_{ii} = (0.9, 0.65, 0.65)$.

The other three are designed to see how sensitive the bias is to relate to the misclassification of each category:

Type 7: $\theta_{ii} = (0.6, 0.9, 0.9)$,

Type 8: $\theta_{ii} = (0.9, 0.6, 0.9)$,

Type 9: $\theta_{ii} = (0.9, 0.9, 0.6)$.

To explain, Type 7 stands for the scenario that category 0 is misclassified most frequently; Type 8 stands for the scenario that category 1 is misclassified most frequently; Type 9 stands for the scenario that category 2 is misclassified most frequently.

In order to calculate exact and approximate biases, we first need to derive $\xi_{d,ab}$ (section 3.1.1), which is equivalent to:

$$\begin{aligned}\xi_{d,ab} &= \frac{P(E=a|Y=d)}{P(E=b|Y=d)} \\ &= \frac{P(E=a,Y=d)}{P(E=b,Y=d)} \\ &= \frac{P(Y=d|E=a)P(E=a)}{P(Y=d|E=b)P(E=b)},\end{aligned}$$

where

$$P(Y=1|E) = \frac{\exp(\beta_0+\beta_1E_1+\beta_2E_2)}{1+\exp(\beta_0+\beta_1E_1+\beta_2E_2)}.$$

Then the exact and first order Taylor approximate asymptotic biases can be easily calculated by plugging $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ into equation (3.4)-(3.11).

The outcomes are demonstrated in Table A.1-A.3 and Table A.6-A.8 based on various settings of $\boldsymbol{\beta}$. There are several findings according to the results:

First, **the length of biases (i.e. $(\Delta_0^2+\Delta_1^2+\Delta_2^2)^{1/2}$) does not necessarily show an increasing trend as the misclassification probabilities increase**. For instance, although the Type 3 case suffers a more severe misclassification problem than the Type 2 case, according to Table A.1, the length of the Type 3 misclassification case is 0.3279, which is smaller than the length of the Type 2 one (i.e. 0.5396).

Second, **both Table A.3 and A.8 demonstrate the bias is most sensitive to misclassification in category 2 when $\boldsymbol{\beta}$ equals (0.1, 0.2, 0.8) and (0.1, -0.2, 0.8), respectively**. That is, the length has the biggest value when the most misclassification occurs in category 2 compared to the other two cases. The results also imply it is most

important to have the category 2 classified in the correct category in order to reduce the biases.

Third, by examine the relative biases, we see **the differences between the exact results and the approximate results could be profound, even not severe misclassification we have**. For instance, the relative biases are about 0.8 for differential misclassification and about 0.6 for non-differential misclassification when least severe misclassification Type 1 and Type 4 are assigned to each case, respectively (see Table A.6 and A.7). **Due the the possible remarkable differences between the exact and approximate results, the exact formulas are generally suggested compared to the approximate ones in order to calculate the asymptotic biases**.

Fourth, in general, **the differences between the exact and approximate results increase as the number of errors increases. When the misclassification problem is serious, there are fundamental differences between these two results**. For instance, in Table A.6, when $\theta_{0,ii} = (0.9, 0.65, 0.65)$ and $\theta_{1,ii} = (0.7, 0.7, 0.7)$ (i.e. Type 3), the length calculated from the approximate results (1.1878) is almost twice the length calculated from the exact ones (0.6705). This conclusion makes sense because the Taylor approximation is derived around the classification matrix I_m, which represents no misclassification of E. It is supposed to have large differences between the exact and approximate outcomes for those scenarios where we have serious misclassification. By looking at it another way, a serious misclassification issue is probably encountered in exposure if we obtain significant differences between the exact and approximate answers.

4.1.2 One Misclassified Exposure and One Perfectly Measured Covariate in the Model

This subsection studies the model including one three-categorical exposure E and one error-free binary covariate Z. I build a model to relate (E,Z) to Y:

$$\text{logit}P(Y = 1|E, Z) = \beta_0 + \beta_1 E_1 + \beta_2 E_2 + \beta_3 Z. \tag{4.3}$$

Likewise, a surrogate model is built to relate (X,Z) to Y:

$$\text{logit}P(Y = 1|X, Z) = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 Z. \tag{4.4}$$

It is common that in reality exposures are correlated. Therefore, let us assume the joint distribution of E and Z is:

	E=0	E=1	E=2
Z=0	0.15	0.05	0.1
Z=1	0.25	0.15	0.3

So that we have the marginal distribution: $P(E = 0) = 0.4$ $P(E = 1) = 0.2$, $P(E = 2) = 0.4$, $P(Z = 0) = 0.3$, and $P(Z = 1) = 0.7$.

In order to check if misclassification in E has an effect on the perfectly measured covariate Z, exact and approximate biases are calculated. The first step is to calculate $\boldsymbol{\xi}$ (section 3.2.1), which can be expressed as:

$$\begin{aligned}
\xi_{d,ab|k} &= \frac{P(Y = d, E = a|Z = k)}{P(Y = d, E = b|Z = k)} \\
&= \frac{P(Y = d, E = a, Z = k)}{P(Y = d, E = b, Z = k)} \\
&= \frac{P(Y = d|E = a, Z = k)P(E = a, Z = k)}{P(Y = d|E = b, Z = k)P(E = b, Z = k)},
\end{aligned}$$

where

$$P(Y = 1|E, Z) = \frac{exp(\beta_0 + \beta_1 E_1 + \beta_2 E_2 + \beta_3 Z)}{1 + exp(\beta_0 + \beta_1 E_1 + \beta_2 E_2 + \beta_3 Z)}.$$

The second step is to calculate the exact and approximate biases by pluging $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ into the expressions (3.15)-(3.26).

The outcomes are illustrated in Table B.1-B.3 and Table B.6-B.8 according to various settings of $\boldsymbol{\beta}$. The results generally agree with those conclusions gained from Table A.1-A.3 and Table A.6-A.8. Besides, there is one additional finding specifically for adding the error-free Z into the model. That is, **misclassification of $\boldsymbol{E}$ has an influence on the coefficient of $\boldsymbol{Z}$ based on the fact that all Δ_3 are unequal to zero**. This statement is also in

agreement with the theoretical results expressed in equation (3.17), (3.20), (3.23) and (3.26).

Particularly, I also calculate the asymptotic biases when E and Z are independent. The results are shown in Table B.11-B.16. The correlation between Y and Z is arranged in an ascending order by increasing β_3 for the differential cases (from B.11 to B.13) and for the non-differential scenarios (from B.14 to B.16). Note that, under the assumption of E and Z independent, $P(E,Z) = P(E) * P(Z)$. Therefore, $\xi_{d,ab|k}$ becomes:

$$\begin{aligned}
\xi_{d,ab|k} &= \frac{P(Y=d|E=a,Z=k)P(E=a,Z=k)}{P(Y=d|E=b,Z=k)P(E=b,Z=k)} \\
&= \frac{P(Y=d|E=a,Z=k)P(E=a)P(Z=k)}{P(Y=d|E=b,Z=k)P(E=b)P(Z=k)} \\
&= \frac{P(Y=d|E=a,Z=k)P(E=a)}{P(Y=d|E=b,Z=k)P(E=b)}.
\end{aligned}$$

The results show **the independency between E and Z does not yield the unbiasedness of** γ_3. In addition, **under the assumption of Z independent of E, the absolute bias for the coefficient of the error-free Z goes up as the correlation between Z and Y increases (i.e.** β_3 **increases)**.

4.2 Simulation Studies

Simulation studies are designed to show how the estimated biases from simulated finite-sample-size datasets approach the asymptotic results from the formulas derived in Chapter 3 as sample size increases. I will compare the asymptotic biases Δ obtained from the example section (section 4.1) with the estimated biases $\widehat{\Delta}$ obtained from the simulation studies of different sample sizes, where $\Delta = (\Delta_0, \Delta_1, \Delta_2)'$ and $\widehat{\Delta} = (\widehat{\Delta_0}, \widehat{\Delta_1}, \widehat{\Delta_2})'$ for without Z case, and $\Delta = (\Delta_0, \Delta_1, \Delta_2, \Delta_3)'$ and $\widehat{\Delta} = (\widehat{\Delta_0}, \widehat{\Delta_1}, \widehat{\Delta_2}, \widehat{\Delta_3})'$ for with Z situation.

4.2.1 Only One Misclassified Exposure in the Model

I give the step-wise guideline for simulation studies when only one misclassified exposure involved (4.1) (Table 4.1). The sample size n of 50, 500 and 5000 and the repetition I of 2000 times (i.e. 2000 iteration) are chosen for the above procedure. Note the choice of the number

Table 4.1: Guideline for simulation studies when E is in the model

Step 1: Generate the true exposure E with sample size n from the multinomial distribution. That is, $E \sim Multinomial(n, 0.4, 0.2, 0.4)$;
Step 2: Generate the outcome Y with size n from the Bernoulli distribution, i.e. $Y \sim Bernoulli(P(Y=1\|E))$ where $P(Y=1\|E) = \frac{exp(\beta_0+\beta_1 E_1+\beta_2 E_2)}{1+exp(\beta_0+\beta_1 E_1+\beta_2 E_2)}$;
Step 3: Calculate $\widehat{\beta}$ by fitting the generated Y and E_j into the logistic regression model (4.1), where E_j are dummy variables of E and $E=0$ is the reference category;
Step 4: Generate the surrogate X with size n from the multinomial distribution. That is, $X \sim Multinomial(n, \theta_{d,0j}, \theta_{d,1j}, \theta_{d,2j})$ for given $Y=d$ and $E=j$;
Step 5: Calculate $\widehat{\gamma}$ by fitting the data (X, Y) into the logistic regression model (4.2);
Step 6: Calculate $\widehat{\beta}$ - $\widehat{\gamma}$.
Step 7: Repeat Step 1-6 I times and compute the average of ($\widehat{\beta}$ - $\widehat{\gamma}$)

of iterations and the sample size depend on the need of the researcher. Ideally, no standard error is along with the average of ($\widehat{\beta}$ - $\widehat{\gamma}$) if I represents an infinite number. An increasing sequence of sample sizes (50, 500, 5000) is employed to verify the validity of the asymptotic bias formulas. Also note the above guideline is described for differential misclassification. For non-differential situation, one can set $\theta_{0,ij} = \theta_{1,ij}$ and then follow the same steps as above.

The numerical results of the above simulation studies are presented in Table A.4, A.5, A.9, and A.10. Here is the summary of findings. First, we see that **the estimated biases approach the asymptotic biases as the sample size increases**. Second, **a remarked difference could arise between Δ_i and $\widehat{\Delta_i}$**. Note that, a relative bias, which equals to $|\frac{\text{asymptotic bias-approximate bias}}{\text{asymptotic bias}}|$, can be calculated as a measure of the difference between exact and estimated results. For instance, the relative bias for Δ_1 is about 3.24 (i.e. $|\frac{0.0187-(-0.0418)}{0.0187}|$) when we have Type 5 misclassification, $\boldsymbol{\beta}$=(0.1,0.2,0.4), and sample size 50 for the simulation study (Table A.5). This result implies a significant difference between Δ_1 and $\widehat{\Delta_1}$ since the difference between Δ_1 and $\widehat{\Delta_1}$ is about three times of the Δ_1 itself. Third, **the differences between the asymptotic and estimated results become bigger when**

misclassification becomes more severe for varied sample size. This conclusion may suggest **a larger sample would be needed for a more severe misclassification scenario as to verify the asymptotic results**.

4.2.2 One Misclassified Exposure and One Perfectly Measured Covariate in the Model

The guideline for simulation studies when model (4.3) is under consideration is shown in Table 4.2. The sample size n of 50, 500 and 5000 and the repetition I of 2000 times (i.e. 2000 iteration) are chosen for the above procedure. Note the choice of the number of iterations and the sample size depend on the need of the researcher. Ideally, no standard error is along with the average of ($\widehat{\beta}$ - $\widehat{\gamma}$) if I represents an infinite number. An increasing sequence of sample sizes (50, 500, 5000) is employed to verify the validity of the asymptotic bias formulas. Also note the above guideline is described for differential misclassification. For non-differential situation, one can set $\theta_{0,ij} = \theta_{1,ij}$ and then follow the same steps as above.

The results of the above simulation studies are illustrated in Table B.4, B.5, B.9, and B.10. The findings are similar to those described in the previous subsection when Z is not in the model. Besides, we can see **although the estimated bias results show a trend of approaching the asymptotic biases, it is not as obvious as the results when *Z* is not involved in the study**. This is due to the fact that discrete data contain less information than continuous data. In addition, adding one more discrete variable Z in the model makes the situation more complex as more information is needed and more parameters are required to be estimated. Therefore, a larger sample size would be necessary in order to verify the asymptotic results.

Table 4.2: Guideline for simulation studies when E and Z are in the model

Step 1: Generate size n of E when $E \sim Multinomial(n, 0.4, 0.2, 0.4)$.
And generate size n of Z when $Z \sim Bernoulli(P(Z = 1\|E = j))$,
where $P(Z = k\|E = j) = P(E = j, Z = k)/P(E = j)$;
Step 2: Generate n Y's following the logistic regression model with:
$P(Y = 1\|E, Z) = \frac{exp(\beta_0+\beta_1 E_1+\beta_2 E_2+\beta_3 Z)}{1+exp(\beta_0+\beta_1 E_1+\beta_2 E_2+\beta_3 Z)}$;
Step 3: Calculate $\widehat{\beta}$ by fitting the generated Y, the indicator variables E_j,
and Z into the logistic regression model (4.3);
Step 4: Generate size n of surrogate X when $X \sim Multinomial(n, \theta_{d,0j}, \theta_{d,1j}, \theta_{d,2j})$
for given $Y = d$ and $E = j$. Note that this step does not depend on Z;
Step 5: Calculate $\widehat{\gamma}$ by fitting the generated Y, the indicator variables X_i,
and Z into the logistic model (4.4);
Step 6: Finally, the estimated biases are ($\widehat{\beta}$ - $\widehat{\gamma}$).
Step 7: Repeat Step 1-6 I times and compute the average of ($\widehat{\beta}$ - $\widehat{\gamma}$)

Chapter 5

Discussion

5.1 Summary

In this thesis, I extend Davidov et al.'s[10] work from a binary case to a multi-categorical scenario with the aim of examining the potential influences of a misclassified exposure on the coefficients of a logistic regression model. I have studied two circumstances: one treats only E as the independent variable and another one includes both E and Z as independent variables, where E is a multi-categorical exposure that is subject to misclassification and Z is an error-free binary covariate. I derive the exact and approximate asymptotic bias formulas for either differential or non-differential misclassification. In practice, differential misclassification would be appropriate for case-control studies while non-differential misclassification would be appropriate for cohort studies. Specifically, the sufficient but not necessary conditions to determine the directions of the biases are explored for the non-differential situation.

An example and simulation study is also provided to evaluate the asymptotic and estimated biases caused by a misclassified three-categorical exposure under various scenarios. Hypothetical examples have been used to examine the magnitude and the directions of the asymptotic biases by comparing the length of the bias vector. I have found that, first, the length of the biases may show a decreasing trend when the magnitude of misclassification increases. Second, my examples show the category 2 is the most sensitive category of E. Because the biggest overall bias (i.e. length) is obtained when I assign most misclassification in category 2. Third, there is an increment in the differences between the exact and approximate biases as misclassification becomes severe. Fourth, when Z is considered in the model, misclassification of E causes a bias in the coefficient of Z regardless of whether E and Z

independent. Simulation studies are also conducted for the finite sample cases. The results show the estimated biases approach the asymptotic biases as sample size increases.

5.2 Significance of My Work

The theoretical findings can provide insightful guide in practice. To explain, the formulas derived in this thesis help researchers assess the magnitude and the directions of the asymptotic biases for large scale studies. If researchers have empirical knowledge from historical studies or similar studies about misclassification rates $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, and marginal distribution of E for model (3.1) or the joint distribution of E and Z for model (3.12), then the asymptotic formulas derived in Chapter 3 can help determine the magnitude of the biases in regression coefficients. For instance, researchers have an internal validation dataset where the surrogate and true information about the exposure E is recoded. Then the marginal distribution of E can be found by using the true data, and the misclassification rates $\boldsymbol{\theta}$ can be figured out by employing both surrogate and true data. In addition, because $exp(\beta_i)$ can be interpreted in terms of odds or odds ratios as shown in section 3.1.2 and 3.2.2, researchers can obtain $exp(\beta_i)$ and β_i based on the validation data. Therefore, $\xi_{d,ab}$ can be computed by using formulas $\xi_{d,ab} = \frac{P(Y=d|E=a)P(E=a)}{P(Y=d|E=b)P(E=b)}$ and $P(Y=1|E) = \frac{\exp(\beta_0+\beta_1 E_1+\ldots+\beta_{m-1}E_{m-1})}{1+\exp(\beta_0+\beta_1 E_1+\ldots+\beta_{m-1}E_{m-1})}$. Finally, the asymptotic biases can be calculated by plugging $\xi_{d,ab}$ and $\boldsymbol{\theta}$ into the derived bias formulas in Chapter 3. If the magnitude of the bias is substantially large, then correction methods that take misclassification into account should be considered.

In addition, the bias derived in regression coefficients can be translated into the bias in the odds ratios. That is, the exponential of Δ_t, $t \neq 0$; is the ratio of OR_{true} to $\mathrm{OR}_{error-prone}$. Mathematically, this relationship can be expressed as:

$$\begin{aligned}\exp(\Delta_t) &= \frac{\mathrm{OR}_{true}}{\mathrm{OR}_{error-prone}} \\ &= \left(\frac{\mathrm{odds}(P(Y=1|E=t))}{\mathrm{odds}(P(Y=1|E=0))}\right)/\left(\frac{\mathrm{odds}(P(Y=1|X=t))}{\mathrm{odds}(P(Y=1|X=0))}\right).\end{aligned}$$

5.3 Future Work

There are several issues remaining to be tackled in the future. First, all bias expressions derived in this thesis are based on the asymptotic theory; however, in reality, a small sample is often of interest. By comparing asymptotic bias results from my derived formulas with estimated bias results from simulation studies, we can see how increasing the sample size influences on the estimated biases and how the estimated biases approach the asymptotic biases as the sample size goes up. However, a more general conclusion about the relationship between limited sample results and asymptotic outcomes needs further investigation. Second, different from Davidov et al.'s[10] work, no interaction term is assumed in my models. Nevertheless, if investigators are interested in misclassification effects on interaction terms, further work needs to be done. Third, the direction of the bias is sometimes the motivation of the study. I have given the sufficient but not necessary conditions to determine the sign of the bias for the non-differential case. However, a general conclusion for the differential case may become a future interest. Last but not least, Davidov et al.[10] also explore the bias formulas when the binary outcome Y has misclassification and when misclassification of E depends on another covariate Z. The more general bias derivation for these situations when E is a multi-categorical variable is still in need of further study.

Bibliography

[1] B. Everitt. *The Cambridge dictionary of statistics*. Cambridge University Press, New York, 2002.

[2] P. Gustafson. *Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments*. Chapman and Hall/CRC Pres, Boca Raton, Florida, 2004.

[3] J.P. Buonaccorsi. *Measurement error: Models, methods, and applications*. Chapman and Hall/CRC, Boca Raton, Florida, 2010.

[4] S.M. Mwalili. *Bayesian and frequentist approaches to correct for misclassification error with applications to caries research*. PhD thesis, Catholic University of Leuven, Leuven, Belgium, 2006.

[5] K.M. Flegal, P.M. Keyl, and F.J. Nieto. Differential misclassification arising from nondifferential errors in exposure measurement. *American Journal of Epidemiology*, 134(10):1233, 1991.

[6] D. Spiegelman, B. Rosner, and R. Logan. Estimation and inference for logistic regression with covariate misclassification and measurement error in main study validation study designs. *Journal of the American Statistical Association*, 95(449):51–61, 2000.

[7] R.J. Carroll, D. Ruppert, and L.A. Stefanski. *Measurement error in nonlinear models. 2nd ed.* Chapman and Hall, New York, 2006.

[8] R. Chu, P. Gustafson, and N. Le. Bayesian adjustment for exposure misclassification in case–control studies. *Statistics in Medicine*, 29(9):994–1003, 2010.

[9] C.D. Drews, W.D. Flanders, and A.S. Kosinski. Use of two data sources to estimate odds ratios in case-control studies. *Epidemiology*, 4(4):327–335, 1993.

[10] O. Davidov, D. Faraggi, and B. Reiser. Misclassification in Logistic Regression with Discrete Covariates. *Biometrical Journal*, 45(5):541–553, 2003.

[11] P. Webb and C. Bain. *Essential epidemiology: An introduction for students and health professionals*. Cambridge University Press, New York, 2011.

[12] B.B. Gerstman. *Epidemiology kept simple: An introduction to traditional and modern epidemiology*. Wiley-Liss, Hoboken, N.J., 2003.

[13] J. Olsen, K. Christensen, J. Murray, and A. Ekbom. *An introduction to epidemiology for health professionals*. Springer Verlag, New York, 2010.

[14] S.D. Walter and L.M. Irwig. Estimation of test error rates, disease prevalence and relative risk from misclassified data: A review. *Journal of Clinical Epidemiology*, 41(9):923–937, 1988.

[15] D.A. Pierce, D.O. Stram, M. Vaeth, and D.W. Schafer. The errors-in-variables problem: Considerations provided by radiation dose-response analyses of the A-bomb survivor data. *Journal of the American Statistical Association*, 87(418):351–359, 1992.

[16] B. Rosner and R. Gore. Measurement error correction in nutritional epidemiology based on individual foods, with application to the relation of diet to breast cancer. *American Journal of Epidemiology*, 154(9):827, 2001.

[17] B. MacMahon and T.F. Pugh. *Epidemiology: Principles and methods*. Little Brown & Co., Boston, 1970.

[18] H. Brenner, V. Arndt, C. Stegmaier, H. Ziegler, and D. Rothenbacher. Is *Helicobacter pylori* infection a necessary condition for noncardia gastric cancer? *American Journal of Epidemiology*, 159(3):252, 2004.

[19] L. Gordis. *Epidemiology. 3rd.* WB Saunders, Philadelphia, PA, 2004.

[20] C.B. Johannes, S.L. Crawford, and J.B. McKinlay. Interviewer effects in a cohort study. *American Journal of Epidemiology*, 146(5):429, 1997.

[21] W. Fuller. *Measurement error models*. John Wiley & Sons, New York, 1987.

[22] P.J. Brown and W.A. Fuller. *Statistical analysis of measurement error models and applications: Proceedings of the AMS-IMS-SIAM joint summer research conference held June 10-16, 1989, with support from the National Science Foundation and the US Army Research Office*, volume 112. American Mathematical Society, 1990.

[23] J. Buzas, L. Stefanski, and T. Tosteson. *Handbook of Epidemiology*, pages 729–765. Springer, Berlin, 2005.

[24] M. Rudemo, D. Ruppert, and J. Streibig. Random effect models in nonlinear regression with applications to bioassay. *Biometrics*, 45(2):349–362, 1989.

[25] M.J. Morrissey and D. Spiegelman. Matrix methods for estimating odds ratios with misclassified exposure data: Extensions and comparisons. *Biometrics*, 55(2):338–344, 1999.

[26] G.J. Prescott and P.H. Garthwaite. Bayesian analysis of misclassified binary data from a matched case-control study with a validation sub-study. *Statistics in Medicine*, 24(3):379–401, 2005.

[27] M.T. Tsuang, M. Tohen, and G.E.P. Zahner. *Textbook in psychiatric epidemiology*. Wiley-Blackwell, Hoboken, NJ, 1995.

[28] I. Bross. Misclassification in 2 by 2 tables. *Biometrics*, 10(4):478–486, 1954.

[29] D.J. Newell. Misclassification in 2 by 2 tables. *Biometrics*, 19(1):187–188, 1963.

[30] K.T. Copeland, H. Checkoway, A.J. Mcmichael, and R.H. Holbrook. Bias due to misclassification in the estimation of relative risk. *American Journal of Epidemiology*, 105(5):488–495, 1977.

[31] S. Greenland. The effect of misclassification in the presence of covariates. *American Journal of Epidemiology*, 112(4):564–569, 1980.

[32] M. Dosemecl, S. Wacholder, and J.H. Lubin. Does nondifferential misclassification of exposure always bias a true effect toward the null value? *American Journal of Epidemiology*, 132(4):746–748, 1990.

[33] NJ Birkett. Effect of nondifferential misclassification on estimates of odds ratios with multiple levels of exposure. *American Journal of Epidemiology*, 136(3):356, 1992.

[34] C.A. Weinberg, D.M. Umbach, and S. Greenland. When will nondifferential misclassification of an exposure preserve the direction of a trend? *American Journal of Epidemiology*, 140(6):565, 1994.

[35] G. Casella and R.L. Berger. *Statistical inference*. Duxbury Press, Pacific Grove, CA, 2001.

[36] M.A. Espeland and S.L. Hui. A general approach to analyzing epidemiologic data that contain misclassification errors. *Biometrics*, 43(4):1001–1012, 1987.

[37] KF Cheng and HM Hsueh. Correcting bias due to misclassification in the estimation of logistic regression models. *Statistics & Probability letters*, 44(3):229–240, 1999.

[38] C.A. Holcroft and D. Spiegelman. Design of validation studies for estimating the odds ratio of exposure–disease relationships when exposure is misclassified. *Biometrics*, 55(4):1193–1201, 1999.

[39] B.A. Barron. The effects of misclassification on the estimation of relative risk. *Biometrics*, 33(2):414–418, 1977.

[40] R.J. Marshall. Validation study methods for estimating proportions and odds ratios with misclassified data. *Journal of Clinical Epidemiology*, 43(9):941–947, 1990.

[41] S. Greenland. Variance estimation for epidemiologic effect estimates under misclassification. *Statistics in Medicine*, 7(7):745–757, 1988.

[42] R.H. Lyles. A note on estimating crude odds ratios in case–control studies with differentially misclassified exposure. *Biometrics*, 58(4):1034–1036, 2002.

[43] A.S. Kosinski and W.D. Flanders. Evaluating the exposure and disease relationship with adjustment for different types of exposure misclassification: A regression approach. *Statistics in Medicine*, 18(20):2795–2808, 1999.

[44] R.H. Lyles and J. Lin. Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting. *Statistics in Medicine*, 29(22):2297–2309, 2010.

[45] J.O. Berger. *Statistical decision theory and Bayesian analysis, 2nd ed.* Springer Verlag, New York, 1985.

[46] P. Gustafson, N.D. Le, and R. Saskin. Case-control analysis with partial knowledge of exposure misclassification probabilities. *Biometrics*, 57(2):598–609, 2001.

[47] H. Chu, Z. Wang, S.R. Cole, and S. Greenland. Sensitivity analysis of misclassification: a graphical and a Bayesian approach. *Annals of Epidemiology*, 16(11):834–841, 2006.

[48] G.J. Prescott and P.H. Garthwaite. A simple Bayesian analysis of misclassified binary data with a validation substudy. *Biometrics*, 58(2):454–458, 2002.

[49] G.J. Prescott and P.H. Garthwaite. A Bayesian approach to prospective binary outcome studies with misclassification in a binary risk factor. *Statistics in Medicine*, 24(22):3463–3477, 2005.

[50] J. Liu, P. Gustafson, N. Cherry, and I. Burstyn. Bayesian analysis of a matched case–control study with expert prior information on both the misclassification of exposure and the exposure–disease association. *Statistics in Medicine*, 28(27):3411–3423, 2009.

[51] H. K*ü*chenhoff, S.M. Mwalili, and E. Lesaffre. A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics*, 62(1):85–96, 2006.

[52] S.J. Reade-Christopher and L.L. Kupper. Effects of exposure misclassification on regression analyses of epidemiologic follow-up study data. *Biometrics*, 47(2):535–548, 1991.

[53] M. Kuroda and G. Zhi. *From Adam Smith to Michael Porter: Evolution of competitiveness theory*, pages 143–151. World Scientific Pub Co Inc, Australia, 2000.

[54] M.A.G. Viana. Bayesian small-sample estimation of misclassified multinomial data. *Biometrics*, 50(1):237–243, 1994.

[55] M. Ruiz, FJ Girón, CJ Pérez, J. Martín, and C. Rojano. A Bayesian model for multinomial sampling with misclassified data. *Journal of Applied Statistics*, 35(4):369–382, 2008.

[56] J. Liu and Y. Liu. Bias analysis for logistic regression with a misclassified multicategorical covariate: Proceedings of Stat 2011 Canada/ IMST 2011-FIM XX held July 1-4, 2011. Unpublished manuscript, 2011.

Appendix A

Simulation Results When There is One Exposure in The Model

Define that $\boldsymbol{\theta}_{d,ii} = (a, b, c)$, i=0,1,2; d=0,1; are diagonal values of the misclassification matrix given $Y = d$ as stated in section 4.1.1.
Define three types of misclassification for differential misclassification.
Type 1: $\boldsymbol{\theta}_{0,ii} = (0.90, 0.90, 0.90)$ and $\boldsymbol{\theta}_{1,ii} = (0.80, 0.80, 0.80)$;
Type 2: $\boldsymbol{\theta}_{0,ii} = (0.90, 0.65, 0.90)$ and $\boldsymbol{\theta}_{1,ii} = (0.70, 0.70, 0.80)$;
Type 3: $\boldsymbol{\theta}_{0,ii} = (0.90, 0.65, 0.65)$ and $\boldsymbol{\theta}_{1,ii} = (0.70, 0.70, 0.70)$.

Define six types of misclassification for non-differential misclassification. Note that $\boldsymbol{\theta}_{d,ii}$ can be simplified as $\boldsymbol{\theta}_{ii}$ under the assumption of non-differential misclassification.
Type 4: $\boldsymbol{\theta}_{ii} = (0.90, 0.90, 0.90)$;
Type 5: $\boldsymbol{\theta}_{ii} = (0.90, 0.65, 0.90)$;
Type 6: $\boldsymbol{\theta}_{ii} = (0.90, 0.65, 0.65)$;
Type 7: $\boldsymbol{\theta}_{ii} = (0.60, 0.90, 0.90)$;
Type 8: $\boldsymbol{\theta}_{ii} = (0.90, 0.60, 0.90)$;
Type 9: $\boldsymbol{\theta}_{ii} = (0.90, 0.90, 0.60)$.

Define that $\Delta_i = \beta_i - \gamma_i$ are asymptotic biases.

Table A.1: Exact and approximate asymptotic biases for differential misclassification parameters and the corresponding length of bias (Δ_0,Δ_1,Δ_2) when $\beta = (0.1, 0.2, 0.4)$ and the misclassification probabilities are arranged in an ascending order. Note that Relative Bias=$|\frac{\text{Exact-Approx}}{\text{Exact}}|$

	Δ_0		Δ_1		Δ_2		$\|\Delta\|$		
$\boldsymbol{\theta}_{d,ii}$	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Relative Bias
Type 1	0.0762	0.0677	-0.2827	-0.3660	0.0163	0.0144	0.0860	0.1387	0.6134
Type 2	0.2240	0.1963	-0.6944	-0.8274	-0.0852	-0.0669	0.5396	0.7276	0.3485
Type 3	0.2240	0.1963	-0.4634	-0.6005	-0.2511	-0.2169	0.3279	0.4462	0.3608

Table A.2: Exact and approximate asymptotic biases for non-differential misclassification parameters and the corresponding length of bias (Δ_0,Δ_1,Δ_2) when $\beta = (0.1, 0.2, 0.4)$ and the misclassification probabilities are arranged in an ascending order. Note that Relative Bias=$|\frac{\text{Exact-Approx}}{\text{Exact}}|$

	Δ_0		Δ_1		Δ_2		$\|\Delta\|$		
$\boldsymbol{\theta}_{ii}$	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Relative Bias
Type 4	-0.0054	-0.0050	0.0063	0.0061	0.0109	0.0101	0.0002	0.0002	0.1223
Type 5	-0.0176	-0.0174	0.0187	0.0185	0.0356	0.0352	0.0019	0.0019	0.0227
Type 6	-0.0176	-0.0174	-0.0453	-0.0797	0.0418	0.0352	0.0041	0.0079	0.9210

Table A.3: Exact and approximate asymptotic biases for non-differential misclassification parameters and the corresponding length of bias (Δ_0,Δ_1,Δ_2) when $\beta = (0.1, 0.2, 0.8)$. Note Type 7 stands for the scenario that category 0 is misclassified most frequently; Type 8 stands for the scenario that category 1 is misclassified most frequently; Type 9 stands for the scenario that category 2 is misclassified most frequently

	Δ_0		Δ_1		Δ_2		$\|\Delta\|$	
$\boldsymbol{\theta}_{ii}$	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx
Type 7	-0.0079	-0.0050	0.0343	0.0551	0.0258	0.0216	0.0019	0.0035
Type 8	-0.0198	-0.0198	-0.0518	-0.0514	0.0854	0.0863	0.0104	0.0105
Type 9	-0.0054	-0.0050	-0.2128	-0.4013	0.0318	0.0216	0.0463	0.1615

Table A.4: Estimated biases for differential misclassification parameters compared with the exact asymptotic biases obtained in Table A.1 when $\beta = (0.1, 0.2, 0.4)$, the misclassification probabilities are arranged in an ascending order, and sample sizes 50, 500 and 5000 are chosen for finite sample estimation

	asymptotic bias			$n = 50$		
$\boldsymbol{\theta}_{d,ii}$	Δ_0	Δ_1	Δ_2	Δ_0	Δ_1	Δ_2
Type 1	0.0762	-0.2827	0.0163	0.0816	-0.2618	0.0077
Type 2	0.2240	-0.6944	-0.0852	0.2495	-0.8569	-0.1090
Type 3	0.2240	-0.4634	-0.2511	0.2495	-0.4502	-0.3314
	$n = 500$			$n = 5000$		
	Δ_0	Δ_1	Δ_2	Δ_0	Δ_1	Δ_2
Type 1	0.0769	-0.2844	0.0156	0.0754	-0.2815	0.0173
Type 2	0.2260	-0.7008	-0.0867	0.2232	-0.6930	-0.0841
Type 3	0.2260	-0.4678	-0.2543	0.2232	-0.4621	-0.2504

Table A.5: Estimated biases for non-differential misclassification parameters compared with the exact asymptotic biases obtained in Table A.2 when $\beta = (0.1, 0.2, 0.4)$, the misclassification probabilities are arranged in an ascending order, and sample sizes 50, 500 and 5000 are chosen for finite sample estimation

	asymptotic bias			$n = 50$		
$\boldsymbol{\theta}_{ii}$	Δ_0	Δ_1	Δ_2	Δ_0	Δ_1	Δ_2
Type 4	-0.0054	0.0063	0.0109	-0.0067	0.0673	0.0033
Type 5	-0.0176	0.0187	0.0356	-0.0207	-0.0418	0.0344
Type 6	-0.0176	-0.0453	0.0418	-0.0207	0.0124	-0.0063
	$n = 500$			$n = 5000$		
	Δ_0	Δ_1	Δ_2	Δ_0	Δ_1	Δ_2
Type 4	-0.0063	0.0093	0.0111	-0.0061	0.0075	0.0118
Type 5	-0.0177	0.0197	0.0358	-0.0184	0.0207	0.0369
Type 6	-0.0177	-0.0464	0.0411	-0.0184	-0.0438	0.0424

Table A.6: Exact and approximate asymptotic biases for differential misclassification parameters and the corresponding length of bias (Δ_0,Δ_1,Δ_2) when $\beta = (0.1, -0.2, 0.4)$ and the misclassification probabilities are arranged in an ascending order. Note that Relative Bias=$|\frac{\text{Exact-Approx}}{\text{Exact}}|$

	Δ_0		Δ_1		Δ_2		$\|\Delta\|$		
$\boldsymbol{\theta}_{d,ii}$	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Relative Bias
Type 1	0.0930	0.0824	-0.4591	-0.6238	0.0161	0.0142	0.2197	0.3961	0.8034
Type 2	0.2608	0.2288	-0.9215	-1.1413	-0.0854	-0.0644	0.9245	1.3591	0.4701
Type 3	0.2608	0.2288	-0.7375	-1.0438	-0.2420	-0.2144	0.6705	1.1878	0.7715

Table A.7: Exact and approximate asymptotic biases for non-differential misclassification parameters and the corresponding length of bias (Δ_0,Δ_1,Δ_2) when $\beta = (0.1, -0.2, 0.4)$ and the misclassification probabilities are arranged in an ascending order. Note that Relative Bias=$|\frac{\text{Exact-Approx}}{\text{Exact}}|$

	Δ_0		Δ_1		Δ_2		$\|\Delta\|$		
$\boldsymbol{\theta}_{ii}$	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Relative Bias
Type 4	0.0054	0.0050	-0.1270	-0.1633	0.0115	0.0107	0.0163	0.0268	0.6466
Type 5	0.0177	0.0175	-0.1682	-0.1758	0.0375	0.0374	0.0300	0.0326	0.0866
Type 6	0.0177	0.0175	-0.3105	-0.4715	0.0561	0.0374	0.0999	0.2240	1.2421

Table A.8: Exact and approximate asymptotic biases for non-differential misclassification parameters and the corresponding length of bias (Δ_0,Δ_1,Δ_2) when $\beta = (0.1, -0.2, 0.8)$. Note Type 7 stands for the scenario that category 0 is misclassified most frequently; Type 8 stands for the scenario that category 1 is misclassified most frequently; Type 9 stands for the scenario that category 2 is misclassified most frequently

	Δ_0		Δ_1		Δ_2		$\|\Delta\|$	
$\boldsymbol{\theta}_{ii}$	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx
Type 7	0.0080	0.0050	-0.1916	-0.3545	0.0228	0.0237	0.0373	0.1262
Type 8	0.0200	0.0200	-0.2490	-0.2493	0.0922	0.0948	0.0709	0.0715
Type 9	0.0054	0.0050	-0.4266	-0.8019	0.0401	0.0237	0.1836	0.6437

Table A.9: Estimated biases for differential misclassification parameters compared with the exact asymptotic biases obtained in Table A.6 when $\beta = (0.1, -0.2, 0.4)$, the misclassification probabilities are arranged in an ascending order, and sample sizes 50, 500 and 5000 are chosen for finite sample estimation

	asymptotic bias			$n = 50$		
$\boldsymbol{\theta}_{d,ii}$	Δ_0	Δ_1	Δ_2	Δ_0	Δ_1	Δ_2
Type 1	0.0930	-0.4591	0.0161	0.0981	-0.5014	0.0093
Type 2	02608	-0.9215	-0.0854	0.2881	-1.0910	-0.1135
Type 3	0.2608	-0.7375	-0.2420	0.2886	-0.7869	-0.3143
	$n = 500$			$n = 5000$		
	Δ_0	Δ_1	Δ_2	Δ_0	Δ_1	Δ_2
Type 1	0.0937	-0.4597	0.0140	0.0927	-0.4604	0.0170
Type 2	0.2614	-0.9244	-0.0850	0.2610	-0.9242	-0.0843
Type 3	0.2614	-0.7389	-0.2432	0.2610	-0.7404	-0.2411

Table A.10: Estimated biases for non-differential misclassification parameters compared with the exact asymptotic biases obtained in Table A.7 when $\beta = (0.1, -0.2, 0.4)$, the misclassification probabilities are arranged in an ascending order, and sample sizes 50, 500 and 5000 are chosen for finite sample estimation

	asymptotic bias			$n = 50$		
$\boldsymbol{\theta}_{ii}$	Δ_0	Δ_1	Δ_2	Δ_0	Δ_1	Δ_2
Type 4	0.0054	-0.1270	0.0115	0.0031	-0.1196	0.0074
Type 5	0.0177	-0.1682	0.0375	0.0170	-0.2230	0.0297
Type 6	0.0177	-0.3105	0.0561	0.0170	-0.3142	0.0118
	$n = 500$			$n = 5000$		
	Δ_0	Δ_1	Δ_2	Δ_0	Δ_1	Δ_2
Type 4	0.0051	-0.1242	0.0099	0.0052	-0.1277	0.0121
Type 5	0.0161	-0.1638	0.0394	0.0178	-0.1703	0.0386
Type 6	0.0161	-0.3087	0.0573	0.0178	-0.3130	0.0568

Appendix B

Simulation Results When There are One Exposure and One Covariate in The Model

Define that $\boldsymbol{\theta}_{d,ii} = (a, b, c)$, i=0,1,2; d=0,1; are diagonal values of the misclassification matrix given $Y = d$ as stated in section 4.1.1.
Define three types of misclassification for differential misclassification.
Type 1: $\boldsymbol{\theta}_{0,ii} = (0.90, 0.90, 0.90)$ and $\boldsymbol{\theta}_{1,ii} = (0.80, 0.80, 0.80)$;
Type 2: $\boldsymbol{\theta}_{0,ii} = (0.90, 0.65, 0.90)$ and $\boldsymbol{\theta}_{1,ii} = (0.70, 0.70, 0.80)$;
Type 3: $\boldsymbol{\theta}_{0,ii} = (0.90, 0.65, 0.65)$ and $\boldsymbol{\theta}_{1,ii} = (0.70, 0.70, 0.70)$.

Define six types of misclassification for non-differential misclassification. Note that $\boldsymbol{\theta}_{d,ii}$ can be simplified as $\boldsymbol{\theta}_{ii}$ under the assumption of non-differential misclassification.
Type 4: $\boldsymbol{\theta}_{ii} = (0.90, 0.90, 0.90)$;
Type 5: $\boldsymbol{\theta}_{ii} = (0.90, 0.65, 0.90)$;
Type 6: $\boldsymbol{\theta}_{ii} = (0.90, 0.65, 0.65)$;
Type 7: $\boldsymbol{\theta}_{ii} = (0.60, 0.90, 0.90)$;
Type 8: $\boldsymbol{\theta}_{ii} = (0.90, 0.60, 0.90)$;
Type 9: $\boldsymbol{\theta}_{ii} = (0.90, 0.90, 0.60)$.

Define that $\Delta_c = \beta_c - \gamma_c$, c=0,1,2,3; are asymptotic biases.

Table B.1: Exact and approximate asymptotic biases for differential misclassification parameters and the length of bias $(\Delta_0, \Delta_1, \Delta_2, \Delta_3)$ when $\beta = (0.1, 0.2, 0.4, 0.5)$ and the misclassification probabilities are arranged in an ascending order. Note that Relative Bias=$|\frac{\text{Exact-Approx}}{\text{Exact}}|$

	Δ_0		Δ_1		Δ_2		Δ_3		$\|\Delta\|$		Relative
$\boldsymbol{\theta}_{d,ii}$	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Bias
Type 1	0.0897	0.0785	-0.3220	-0.4479	0.0029	0.0036	-0.0202	-0.0163	0.1122	0.2071	0.8461
Type 2	0.2325	0.1975	-0.7530	-0.9912	-0.0937	-0.0681	-0.0127	-0.0019	0.6300	1.0261	0.6289
Type 3	0.2325	0.1975	-0.5348	-0.7643	-0.2596	-0.2181	-0.0127	-0.0019	0.4076	0.6708	0.6457

Table B.2: Exact and approximate asymptotic biases for non-differential misclassification parameters and the length of bias $(\Delta_0, \Delta_1, \Delta_2, \Delta_3)$ when $\beta = (0.1, 0.2, 0.4, 0.5)$ and the misclassification probabilities are arranged in an ascending order. Note that Relative Bias=$|\frac{\text{Exact-Approx}}{\text{Exact}}|$

	Δ_0		Δ_1		Δ_2		Δ_3		$\|\Delta\|$		Relative
$\boldsymbol{\theta}_{d,ii}$	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Bias
Type 4	-0.0036	-0.0033	0.0189	0.0247	0.0091	0.0084	-0.0026	-0.0025	0.0005	0.0007	0.5200
Type 5	-0.0121	-0.0116	0.0307	0.0330	0.0301	0.0294	-0.0082	-0.0088	0.0021	0.0022	0.0500
Type 6	-0.0121	-0.0116	-0.0346	-0.0652	0.0362	0.0294	-0.0082	-0.0088	0.0027	0.0053	0.9553

Table B.3: Exact and approximate asymptotic biases for non-differential misclassification parameters and the length of bias $(\Delta_0, \Delta_1, \Delta_2, \Delta_3)$ when $\beta = (0.1, 0.2, 0.8, 0.5)$. Note Type 7 stands for the scenario that category 0 is misclassified most frequently; Type 8 stands for the scenario that category 1 is misclassified most frequently; Type 9 stands for the scenario that category 2 is misclassified most frequently

	Δ_0		Δ_1		Δ_2		Δ_3		$\|\Delta\|$	
$\boldsymbol{\theta}_{d,ii}$	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx
Type 7	-0.0054	-0.0033	0.0621	0.1344	0.0232	0.0199	-0.0039	-0.0025	0.0044	0.0185
Type 8	-0.0137	-0.0132	-0.0328	-0.0378	0.0792	0.0796	-0.0092	-0.0100	0.0076	0.0080
Type 9	-0.0036	-0.0033	-0.1928	-0.3827	0.0300	0.0199	-0.0026	-0.0025	0.0381	0.1469

Table B.4: Estimated biases for differential misclassification parameters compared with the exact asymptotic biases obtained in Table B.1 when $\beta = (0.1, 0.2, 0.4, 0.5)$, the misclassification probabilities are arranged in an ascending order, and sample sizes 50, 500 and 5000 are chosen for finite sample estimation

	asymptotic bias				n = 50			
$\boldsymbol{\theta}_{d,ii}$	Δ_0	Δ_1	Δ_2	Δ_3	Δ_0	Δ_1	Δ_2	Δ_3
Type 1	0.0897	-0.3220	0.0029	-0.0202	0.0843	-0.0088	0.0087	0.0003
Type 2	0.2325	-0.7530	-0.0937	-0.0127	0.2532	-0.6445	-0.0717	0.0061
Type 3	0.2325	-0.5348	-0.2596	-0.0127	0.2466	-0.1876	-0.3833	0.0169
	n = 500				n = 5000			
	Δ_0	Δ_1	Δ_2	Δ_3	Δ_0	Δ_1	Δ_2	Δ_3
Type 1	0.0770	-0.2858	0.0152	0.0018	0.0764	-0.2841	0.0142	0.0017
Type 2	0.2206	-0.7087	-0.0875	0.0121	0.2170	-0.6989	-0.0867	0.0121
Type 3	0.2146	-0.4770	-0.2578	0.0218	0.2111	-0.4721	-0.2528	0.0215

Table B.5: Estimated biases for non-differential misclassification parameters compared with the exact asymptotic biases obtained in Table B.2 when $\beta = (0.1, 0.2, 0.4, 0.5)$, the misclassification probabilities are arranged in an ascending order, and sample sizes 50, 500 and 5000 are chosen for finite sample estimation

	asymptotic bias				n = 50			
$\boldsymbol{\theta}_{ii}$	Δ_0	Δ_1	Δ_2	Δ_3	Δ_0	Δ_1	Δ_2	Δ_3
Type 4	-0.0036	0.0189	0.0091	-0.0026	0.0016	0.2980	0.0051	-0.0044
Type 5	-0.0121	0.0307	0.0301	-0.0082	-0.0004	0.1646	0.0675	-0.0090
Type 6	-0.0121	-0.0346	0.0362	-0.0082	0.0027	0.2748	-0.0564	-0.0135
	n = 500				n = 5000			
	Δ_0	Δ_1	Δ_2	Δ_3	Δ_0	Δ_1	Δ_2	Δ_3
Type 4	-0.0038	0.0136	0.0145	-0.0062	-0.0016	0.0081	0.0109	-0.0057
Type 5	-0.0134	0.0233	0.0388	-0.0085	-0.0118	0.0205	0.0358	-0.0082
Type 6	-0.0126	-0.0402	0.0442	-0.0099	-0.0111	-0.0437	0.0424	-0.0095

Table B.6: Exact and approximate asymptotic biases for differential misclassification parameters and the length of bias $(\Delta_0, \Delta_1, \Delta_2, \Delta_3)$ when $\beta = (0.1, -0.2, 0.4, 0.5)$ and the misclassification probabilities are arranged in an ascending order. Note that Relative Bias=$|\frac{\text{Exact-Approx}}{\text{Exact}}|$

	Δ_0		Δ_1		Δ_2		Δ_3		$\|\Delta\|$		Relative
$\boldsymbol{\theta}_{d,ii}$	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Bias
Type 1	0.1010	0.0883	-0.5137	-0.7602	0.0081	0.0083	-0.0134	-0.0099	0.2744	0.5859	1.1354
Type 2	0.2579	0.2192	-0.9916	-1.3728	-0.0824	-0.0548	0.0049	0.0162	1.0566	1.9358	0.8321
Type 3	0.2579	0.2192	-0.8124	-1.2753	-0.2390	-0.2048	0.0049	0.0162	0.7836	1.7165	1.1904

Table B.7: Exact and approximate asymptotic biases for non-differential misclassification parameters and the length of bias $(\Delta_0, \Delta_1, \Delta_2, \Delta_3)$ when $\beta = (0.1, -0.2, 0.4, 0.5)$ and the misclassification probabilities are arranged in an ascending order. Note that Relative Bias=$|\frac{\text{Exact-Approx}}{\text{Exact}}|$

	Δ_0		Δ_1		Δ_2		Δ_3		$\|\Delta\|$		Relative
$\boldsymbol{\theta}_{d,ii}$	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Bias
Type 4	0.0036	0.0033	-0.1308	-0.1817	0.0133	0.0123	0.0030	0.0028	0.0173	0.0332	0.9172
Type 5	0.0122	0.0117	-0.1670	-0.1900	0.0430	0.0432	0.0092	0.0099	0.0300	0.0382	0.2751
Type 6	0.0122	0.0117	-0.2994	-0.4856	0.0616	0.0432	0.0092	0.0099	0.0936	0.2380	1.5412

Table B.8: Exact and approximate asymptotic biases for non-differential misclassification parameters and the length of bias $(\Delta_0, \Delta_1, \Delta_2, \Delta_3)$ when $\beta = (0.1, -0.2, 0.8, 0.5)$. Note Type 7 stands for the scenario that category 0 is misclassified most frequently; Type 8 stands for the scenario that category 1 is misclassified most frequently; Type 9 stands for the scenario that category 2 is misclassified most frequently

	Δ_0		Δ_1		Δ_2		Δ_3		$\|\Delta\|$	
$\boldsymbol{\theta}_{d,ii}$	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx
Type 7	0.0054	0.0033	-0.1919	-0.4330	0.0254	0.0254	0.0043	0.0028	0.0375	0.1881
Type 8	0.0138	0.0134	-0.2401	-0.2627	0.0984	0.1015	0.0103	0.0113	0.0676	0.0796
Type 9	0.0036	0.0033	-0.4136	-0.8203	0.0418	0.0254	0.0030	0.0028	0.1729	0.6736

Table B.9: Estimated biases for differential misclassification parameters compared with the exact asymptotic biases obtained in Table B.6 when $\beta = (0.1, -0.2, 0.4, 0.5)$, the misclassification probabilities are arranged in an ascending order, and sample sizes 50, 500 and 5000 are chosen for finite sample estimation

	asymptotic bias				$n = 50$			
$\boldsymbol{\theta}_{d,ii}$	Δ_0	Δ_1	Δ_2	Δ_3	Δ_0	Δ_1	Δ_2	Δ_3
Type 1	0.1010	-0.5137	0.0081	-0.0134	0.0809	-0.4234	0.0103	0.0126
Type 2	0.2579	-0.9916	-0.0824	0.0049	0.2725	-1.0668	-0.0783	0.0289
Type 3	0.2579	-0.8124	-0.2390	0.0049	0.2751	-0.7170	-0.3404	0.0293
	$n = 500$				$n = 5000$			
	Δ_0	Δ_1	Δ_2	Δ_3	Δ_0	Δ_1	Δ_2	Δ_3
Type 1	0.0835	-0.4550	0.0137	0.0173	0.0827	-0.4522	0.0134	0.0173
Type 2	0.2431	-0.9269	-0.0868	0.0337	0.2400	-0.9190	-0.0853	0.0337
Type 3	0.2386	-0.7461	-0.2462	0.0409	0.2357	-0.7427	-0.2404	0.0405

Table B.10: Estimated biases for non-differential misclassification parameters compared with the exact asymptotic biases obtained in Table B.7 when $\beta = (0.1, -0.2, 0.4, 0.5)$, the misclassification probabilities are arranged in an ascending order, and sample sizes 50, 500 and 5000 are chosen for finite sample estimation

	asymptotic bias				$n = 50$			
$\boldsymbol{\theta}_{ii}$	Δ_0	Δ_1	Δ_2	Δ_3	Δ_0	Δ_1	Δ_2	Δ_3
Type 4	0.0036	-0.1308	0.0133	0.0030	-0.0036	-0.0774	0.0099	0.0015
Type 5	0.0122	-0.1670	0.0430	0.0092	0.0229	-0.2260	0.0625	0.0052
Type 6	0.0122	-0.2994	0.0616	0.0092	0.0333	-0.2459	-0.0101	-0.0066
	$n = 500$				$n = 5000$			
	Δ_0	Δ_1	Δ_2	Δ_3	Δ_0	Δ_1	Δ_2	Δ_3
Type 4	0.0011	-0.1206	0.0142	0.0043	0.0027	-0.1248	0.0113	0.0051
Type 5	0.0115	-0.1621	0.0418	0.0074	0.0133	-0.1666	0.0394	0.0079
Type 6	0.0128	-0.3022	0.0604	0.0054	0.0146	-0.3067	0.0596	0.0058

Table B.11: With the assumption of Z independent of E, exact and approximate asymptotic biases for differential misclassification parameters and the length of bias $(\Delta_0, \Delta_1, \Delta_2, \Delta_3)$ when β = $(0.1, 0.2, 0.4, 0.5)$ and the misclassification probabilities are arranged in an ascending order

	Δ_0		Δ_1		Δ_2		Δ_3		$\|\|\Delta\|\|$	
$\boldsymbol{\theta}_{d,ii}$	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx
Type 1	0.0762	0.0677	-0.2827	-0.3660	0.0163	0.0144	0.0009	0.0008	0.0860	0.1387
Type 2	0.2240	0.1963	-0.6944	-0.8274	-0.0852	-0.0669	0.0006	0.0001	0.5396	0.7276
Type 3	0.2240	0.1963	-0.4634	-0.6005	-0.2511	-0.2169	0.0006	0.0001	0.3279	0.4462

Table B.12: With the assumption of Z independent of E, exact and approximate asymptotic biases for differential misclassification parameters and the length of bias $(\Delta_0, \Delta_1, \Delta_2, \Delta_3)$ when β = $(0.1, 0.2, 0.4, 1)$ and the misclassification probabilities are arranged in an ascending order

	Δ_0		Δ_1		Δ_2		Δ_3		$\|\|\Delta\|\|$	
$\boldsymbol{\theta}_{d,ii}$	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx
Type 1	0.0762	0.0677	-0.2827	-0.3660	0.0163	0.0144	0.0017	0.0014	0.0860	0.1387
Type 2	0.2240	0.1963	-0.6944	-0.8274	-0.0852	-0.0669	0.0011	0.0002	0.5396	0.7276
Type 3	0.2240	0.1963	-0.4634	-0.6005	-0.2511	-0.2169	0.0011	0.0002	0.3279	0.4462

Table B.13: With the assumption of Z independent of E, exact and approximate asymptotic biases for differential misclassification parameters and the length of bias $(\Delta_0, \Delta_1, \Delta_2, \Delta_3)$ when β = $(0.1, 0.2, 0.4, 3)$ and the misclassification probabilities are arranged in an ascending order

	Δ_0		Δ_1		Δ_2		Δ_3		$\|\|\Delta\|\|$	
$\boldsymbol{\theta}_{d,ii}$	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx
Type 1	0.0762	0.0677	-0.2827	-0.3660	0.0163	0.0144	0.0031	0.0025	0.0860	0.1387
Type 2	0.2240	0.1963	-0.6944	-0.8274	-0.0852	-0.0669	0.0020	0.0003	0.5396	0.7276
Type 3	0.2240	0.1963	-0.4634	-0.6005	-0.2511	-0.2169	0.0020	0.0003	0.3279	0.4462

Table B.14: With the assumption of Z independent of E, exact and approximate asymptotic biases for non-differential misclassification parameters and the length of bias $(\Delta_0, \Delta_1, \Delta_2, \Delta_3)$ when $\beta = (0.1, 0.2, 0.4, 0.5)$ and the misclassification probabilities are arranged in an ascending order

	Δ_0		Δ_1		Δ_2		Δ_3		$\|\|\Delta\|\|$	
$\boldsymbol{\theta}_{d,ii}$	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx
Type 4	-0.0054	-0.0050	0.0063	0.0061	0.0109	0.0101	0.0001	0.0001	0.0002	0.0002
Type 5	-0.0176	-0.0174	0.0187	0.0185	0.0356	0.0352	0.0004	0.0004	0.0019	0.0019
Type 6	-0.0176	-0.0174	-0.0453	-0.0797	0.0418	0.0352	0.0004	0.0004	0.0041	0.0079

Table B.15: With the assumption of Z independent of E, exact and approximate asymptotic biases for non-differential misclassification parameters and the length of bias $(\Delta_0, \Delta_1, \Delta_2, \Delta_3)$ when $\beta = (0.1, 0.2, 0.4, 1)$ and the misclassification probabilities are arranged in an ascending order

	Δ_0		Δ_1		Δ_2		Δ_3		$\|\|\Delta\|\|$	
$\boldsymbol{\theta}_{d,ii}$	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx
Type 4	-0.0054	-0.0050	0.0063	0.0061	0.0109	0.0101	0.0002	0.0002	0.0002	0.0002
Type 5	-0.0176	-0.0174	0.0187	0.0185	0.0356	0.0352	0.0007	0.0007	0.0019	0.0019
Type 6	-0.0176	-0.0174	-0.0453	-0.0797	0.0418	0.0352	0.0007	0.0007	0.0041	0.0079

Table B.16: With the assumption of Z independent of E, exact and approximate asymptotic biases for non-differential misclassification parameters and the length of bias $(\Delta_0, \Delta_1, \Delta_2, \Delta_3)$ when $\beta = (0.1, 0.2, 0.4, 3)$ and the misclassification probabilities are arranged in an ascending order

	Δ_0		Δ_1		Δ_2		Δ_3		$\|\|\Delta\|\|$	
$\boldsymbol{\theta}_{d,ii}$	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx
Type 4	-0.0054	-0.0050	0.0063	0.0061	0.0109	0.0101	0.0004	0.0004	0.0002	0.0002
Type 5	-0.0176	-0.0174	0.0187	0.0185	0.0356	0.0352	0.0013	0.0014	0.0019	0.0019
Type 6	-0.0176	-0.0174	-0.0453	-0.0797	0.0418	0.0352	0.0013	0.0014	0.0041	0.0079